Selected Works in Probability and Statistics

Jianqing Fan
Ya'acov Ritov
C.F. Jeff Wu *Editors*

Selected Works of Peter J. Bickel

Springer

Selected Works in Probability and Statistics

For further volumes:
http://www.springer.com/series/8556

Jianqing Fan • Ya'acov Ritov • C.F. Jeff Wu
Editors

Selected Works of Peter J. Bickel

Springer

Editors
Jianqing Fan
Department of Operations Research
and Financial Engineering
Princeton University
Princeton
New Jersey, USA

Ya'acov Ritov
Department of Statistics
Hebrew University of Jerusalem
Jerusalem, Israel

C.F. Jeff Wu
School of Industrial and Systems
Engineering
Georgia Institute of Technology
Atlanta, USA

ISBN 978-1-4614-5543-1 ISBN 978-1-4614-5544-8 (eBook)
DOI 10.1007/978-1-4614-5544-8
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012948852

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

I am very grateful to Jianqing, Yanki, and Jeff for organizing this collection of high points in my wanderings through probability theory and statistics, and to the friends and colleagues who commented on some of these works, and without whose collaborations many of these papers would not exist.

Statistics has contacts with, contributes to, and draws from so many fields that there is a nearly infinite number of questions that arise, ranging from those close to particular applications to ones that are at a distance and essentially mathematical. As these papers indicate I've enjoyed all types and have believed in the mantra that ideas developed for solving one problem may unexpectedly prove helpful in very different contexts. The field has, under the pressure of massive, complex, high dimensional data, moved beyond the paradigms established by Fisher, Neyman, and Wald long ago. Despite my unexpectedly advanced age I find it to be so much fun that I won't quit till I have to.

Berkeley, California, USA Peter J. Bickel

Preface

Our civilization depends largely on our ability to record major historical events, such as philosophical thoughts, scientific discoveries, and technological inventions. We manage these records through the collection, organization, presentation, and analysis of past events. This allows us to pass our knowledge down to future generations, to let them learn how our ancestors dealt with similar situations that led to the outcomes we see today. The history of statistics is no exception. Despite its long history of applications that have improved social wellbeing, systematic studies of statistics to understand random phenomena are no more than a century old. Many of our professional giants have devoted their lives to expanding the frontiers of statistics. It is of paramount importance for us to record their discoveries, to understand the environments under which these discoveries were made, and to assess their impacts on shaping the course of development in the statistical world. It is with this background that we enthusiastically edit this volume.

Since obtaining his Ph.D. degree at the age of 22, Peter Bickel's 50 years of distinguished work spans the revolution of scientific computing and data collection, from vacuum tubes for processing and small experimental data to today's supercomputing and automated massive data scanning. The evolution of scientific computing and data collection has a profound impact on statistical thinking, methodological developments, and theoretical studies, thus creating evolving frontiers of statistics.

Peter Bickel has been a leading figure at the forefront of statistical innovations. His career encompasses the majority of statistical developments in the last half-century, which is about half of the entire history of the systematic development of statistics. We therefore select some of his major papers at the frontiers of statistics and reprint them here along with comments on their novelty and importance at that time and their impacts on the subsequent development. We hope that this will enable future generations of statisticians to gain some insights on these exciting statistical developments, help them understand the environment under which this research was conducted, and inspire them to conduct their own research to address future problems.

Peter Bickel's research began with his thesis work on multivariate analysis under the supervision of Erich Lehmann, followed by his work on robust statistics, semiparametric and nonparametric statistics, and present work on high-dimensional statistics. His work demonstrates the evolution of statistics over the last half-century, from classical finite dimensional data in the 1960s and 1970s, to moderate-dimensional data in the 1980s and 1990s, and to high-dimensional data in the first decade of this century. His work exemplifies the idea that statistics as a discipline grows stronger when it confronts the important problems of great social impact while providing a fundamental understanding of these problems and their associated methods that push forward theory, methodology, computation, and applications. Because of the varied nature of Bickel's work, it is a challenge to select his papers for this volume. To help readers understand his contributions from a historical prospective, we have divided his work into the following eight areas: "Rank-based nonparametric statistics", "Robust statistics", "Asymptotic theory", "Nonparametric function estimation", "Adaptive and efficient estimation", "Bootstrap and resampling methods", "High-dimensional statistical learning", and "Miscellaneous". The division is imperfect and somewhat artificial. The work of a single paper can impact the development of multiple areas. We acknowledge that omissions and negligence are inevitable, but we hope to give readers a broad view on Bickel's contributions.

This volume includes new photos of Peter Bickel, his biography, publication list, and a list of his students. We hope this will give the readers a more complete picture of Peter Bickel, as a teacher, a friend, a colleague, and a family man. We include a short foreword by Peter Bickel in this volume.

We are honored to have the opportunity to edit this Selected Work of Peter Bickel and to present his work to the readers. We are grateful to Peter Bühlmann, Peter Hall, Hans-Georg Müeller, Qiman Shao, Jon Wellner, and Willem van Zwet for their dedicated contributions to this volume. Without their in-depth comments and prospects, this volume would not have been possible. We are grateful to Nancy Bickel for her encouragement and support of this project, including the supply of a majority of photos in this book. We would also like to acknowledge Weijie Gu, Nina Guo, Yijie Dylan Wang, Matthias Tan and Rui Tuo for their help in typing some of the comments, collecting of Bickel's bibliography and list of students, and typesetting the whole book. We are indebted to them for their hard work and dedication. We would also like to thank Marc Strauss, Senior Editor, Springer Science and Business Media, for his patience and assistance.

Princeton, NJ, USA — Jianqing Fan
Jerusalem, Israel — Ya'acov Ritov
Atlanta, GA, USA — C.F. Jeff Wu

Contents

Biography of Peter J. Bickel

Peter John Bickel was born Sept. 21, 1940, in Bucharest, Romania, to a Jewish family. His father, Eliezer Bickel, was a medical doctor, researcher and philosopher. His mother, Madeleine, ran the household. After World War II, the family left Romania for Paris in 1948, and moved to Toronto in 1949. His father died in 1951 when he was eleven. He moved to California with his mother in 1957, having finished 5 years of high school in Ontario. He started his undergraduate study at Caltech in 1957 but only stayed for 2 years before transferring to the University of California, Berkeley in 1959. For some inexplicable reason, a substantial number of leading statisticians of his generation came from Caltech, where statistics was not taught. They include Larry Brown, Brad Efron, Carl Morris, and Chuck Stone, among others. At Berkeley he obtained his bachelor's degree in mathematics in 1 year. After quickly obtaining a Master's degree in mathematics, he started his doctoral study in 1961 in the Statistics Department. He obtained his Ph.D. degree in 1963 at the age of 22 under the supervision of Erich Lehmann. He and Lehmann later became close friends. He was immediately hired by the Department, which marked the beginning of his long association with and loyalty to Berkeley. He served as Chair of the Statistics Department (twice) and Dean of the Physical Sciences (twice). He officially retired from Berkeley in 2006 but has continued to maintain his office and an active research program in the Department.

When Bickel joined the Berkeley Statistics Department in the early 1960s, it boasted some of the leading figures in the statistics profession: Jerzy Neyman (its founder), David Blackwell, Joe Hodges, Lucien LeCam, Erich Lehmann, Michel Loeve, and Henry Scheffe, among others. During his student days, he met Kjell Doksum and Yossi Yahav who became close friends and collaborators. He coauthored a widely used textbook (Bickel and Doksum 2001) in mathematical statistics with Doksum. He made several visits to Israel to collaborate with Yahav, including his sabbatical in 1981 in Jerusalem, when Yahav introduced him to a graduate student named Ya'acov Ritov. Bickel became Ritov's chief thesis advisor. They have subsequently collaborated on many papers for the next 30 years. Among Bickel's coauthors, Ritov has the unique honor of having written the most papers with him. Another of his long term collaborators and close friends is Willem van

Zwet from the University of Leiden. They met briefly in the 1960s but started working together in asymptotic theory when van Zwet visited Berkeley in 1972. On the personal side, he married Nancy Kramer in 1964; they had two children Amanda and Stephen and five grandchildren. His attachment to his children and grandchildren has influenced his latest choices of research in weather prediction and genomics, since his daughter Amanda lives in Boulder and his son Stephen outside Washington, D.C. (Ritov 2011). He and Nancy have enjoyed a "loving and intellectually lively family life".

Bickel has made wide-ranging contributions to statistical science. As his students, each of us had just a glimpse of the total picture. Only during the compilation of this volume, did we begin to comprehend the breadth of his research and the magnitude of his impact. It did not take long for us to realize that, in order to include the necessary in-depth discussions, we would have to divide the collection of papers in this volume into eight categories. The readers may consult another review (Doksum and Ritov 2006) of his research contributions. His research in the early period was mostly theoretical, including rank-based nonparametrics, classical asymptotic theory, robust statistics, higher order asymptotics, and nonparametric function estimation. His ability, at a young age, to pursue serious work in a broad range of areas is unusual. However, he did not shy away from doing applied work. In a 1975 *Science* paper (Bickel et al. 1975), he and coauthors gave an explanation of an apparent gender bias in graduate admissions at UC Berkeley by relating it to Simpson's paradox. Over the years he has continued to expand his research horizon into other areas such as bootstrap/resampling, semiparametric and nonparametric estimation, high dimensional statistics and statistical learning. During this period, his work and impact have grown beyond theoretical statistics. He once said that as he got older, he "became bolder in starting to think seriously about the interaction between theory and applications, - -" (Ritov 2011). His interest in real world applications is evident in his major work in molecular biology, traffic analysis, and weather prediction. The breadth and impact of his work is also reflected in the 60 Ph.D. students (list in this volume) he has supervised so far. The dissertation topics of these 60 students are as varied as one can imagine. He is known to be an effective, helpful and supportive thesis advisor.

For the depth, breadth and impact of his work, Bickel is widely viewed as one of the greatest statisticians and a leading light of his time. He has received many distinguished awards and honors. Only a few are mentioned here. He was the Wald Lecturer and Rietz Lecturer of the IMS and the first recipient of the COPSS Presidents' Award. He received a MacArthur Fellowship, was elected to the National Academy of Sciences, the American Academy of Arts and Sciences, and the Royal Netherlands Academy of Arts and Sciences. He has also received an honorary doctoral degree from the Hebrew University of Jerusalem and was appointed Commander in the Order of Oranje-Nassau by Queen Beatrix of the Netherlands. Among his doctoral students, three have received the COPSS Presidents' Award, which must be a record for a thesis advisor. In spite of the fame and recognition he has received since early days, he remains a very modest person. As his former students, we were surprised to read a statement like “I became more self-confident (after getting the MacArthur Fellowship)” (Ritov 2011).

Besides his busy research, he has rendered dedicated service to the profession and the country. He was the President of the Institute of Mathematical Statistics (IMS) and of the Bernoulli Society. He has served on many national committees and commissions, including those in the National Academy of Sciences, National Research Council, the American Association for the Advancement of Science, and EURANDOM.

While most people at his age either decelerate or become idle, he has maintained a vigorous research program and started working in some new directions in biology and computer science. Some may even claim that since his retirement, he has become more active than before. He once confided to one of us that, without the bounds of official duties, he can now choose the course he wants to teach, and go to the meetings he feels comfortable attending. He seems to enjoy the freedom from his retirement and has found more energy for research “despite his unexpectedly advanced age” (Bickel this volume). In a decade or two from now, we will need to undertake a major update of his career and research.

References

Bickel PJ In: Foreword to selected works this volume

Bickel PJ, Doksum KA (2001) Mathematical statistic: basic ideas and selected topics, vol 1, 2nd edn. Prentice Hall, Upper Saddle River

Bickel PJ, Hammel EA, O’Connell JW (1975) Sex bias in graduate admissions: data from Berkeley. Science 187:398–404

Doksum KA, Ritov Y (2006) Our steps on the Bickel way. In: Fan J, Koul HL (eds) Frontier in statistics: dedicated to Peter John Bickel in Honor of his 65th birthday. Imperial College Press, London, pp 1–10

Ritov Y (2011) A random walk with drift: interview with Peter J. Bickel. Stat Sci 26:155

Publications by Peter J. Bickel

- **Publications**

1. Bickel PJ (1964) On some alternative estimates for shift in the p-variate one sample problem. Ann Math Stat 35:1079–1090
2. Bickel PJ (1965) On some asymptotically nonparametric competitors of Hotelling's T^2. Ann Math Stat 36:160–173
3. Bickel PJ (1965) On some robust estimates of location. Ann Math Stat 36:847–859
4. Bickel PJ, Yahav JA (1965) Renewal theory in the plane. Ann Math Stat 36:946–955
5. Bickel PJ, Yahav JA (1965) The number of visits of vector walks to bounded regions. Isr J Math 3:181–186
6. Bickel PJ, Yahav JA (1966) Asymptotically pointwise optimal procedures in sequential analysis. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. I. University of California Press, Berkeley, pp 401–415
7. Bickel PJ (1966) Some contributions to the theory of order statistics. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. I. University of California Press, Berkeley, pp 575–593
8. Bickel PJ, Hodges JL Jr (1967) The asymptotic theory of Galton's test and a related simple estimate of location. Ann Math Stat 38:73–89
9. Bickel PJ, Blackwell D (1967) A note on Bayes estimates. Ann Math Stat 38:1907–1912
10. Bickel PJ, Yahav JA (1968) Asymptotically optimal Bayes and minimax procedures in sequential estimation. Ann Math Stat 39:442–456
11. Bahadur RR, Bickel PJ (1968) Substitution in conditional expectation. Ann Math Stat 39:377–378
12. Berk RH, Bickel PJ (1968) On invariance and almost invariance. Ann Math Stat 39:1573–1577
13. Bickel PJ, Yahav JA (1969) Some contributions to the asymptotic theory of Bayes solutions. Z. Wahrscheinlichkeitstheorie und verw Geb 11:257–276

14. Bickel PJ (1969) A distribution free version of the Smirnov two sample test in the p-variate case. Ann Math Stat 40:1–23
15. Bickel PJ, Yahav JA (1969) On an A.P.O. rule in sequential estimation with quadratic loss. Ann Math Stat 40:417–426
16. Bickel PJ, Doksum KA (1969) Tests for monotone failure rate based on normalized sample spacings. Ann Math Stat 40:1216–1235
17. Bickel PJ (1969) A remark on the Kolmogorov-Petrovskii criterion. Ann Math Statist 40:1086–1090
18. Bickel PJ (1969) Test for monotone failure rate II. Ann Math Stat 40:1250–1260
19. Bickel PJ (1969) Review of "Theory of Rank Tests" by J. Hajek. J Am Stat Assoc 64:397–399
20. Bickel PJ, Lehmann E (1969) Unbiased estimation in convex families. Ann Math Stat 40:1523–1535
21. Bickel PJ (1969) Une generalisation de type Hajek-Renyi d'une inegalite de M.P. Levy. Resume of No 18 CR Acad Sci Paris 269:713–714
22. Bickel PJ (1970) A Hajek-Renyi extension of Levy's inequality and its applications. Acta Math Acad Sci Hung 21:199–206
23. Bickel PJ, Wichura M (1971) Convergence criteria for multiparameter stochastic processes and some applications. Ann Math Stat 42:1656–1669
24. Bickel PJ (1971) On some analogues to linear combinations of order statistics in the linear model. Stat Decis Theory Relat Top 207–216
25. Bickel PJ (1971) Mathematical statistics, part I, preliminary edition. Holden-Day, San Francisco
26. Bahadur RR, Bickel PJ (1971) On conditional test levels in large samples. Roy memorial volume. University of North Carolina Press
27. Bickel PJ, Yahav JA (1972) On the Wiener process approximation to Bayesian sequential testing problems. In: Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, vol 1. University of California Press, Berkeley, pp 57–83
28. Andrews D, Bickel PJ, Hampel F, Huber PJ, Rogers WH, Tukey JW (1972) Robust estimation of location: survey and advances. Princeton
29. Bickel PJ (1973) On some analogues to linear combinations of order statistics in the linear model. Ann Stat 1:597–616
30. Bickel PJ (1973) On the asymptotic shape of Bayesian sequential tests of $\theta \leq 0$ versus $\theta > 0$ for exponential families. Ann Stat 1:231–240
31. Bickel PJ, Rosenblatt M (1973) On some global measures of the deviations of density function estimates. Ann Stat 1:1071–1095
32. Bickel PJ, Rosenblatt M (1973) Stationary random fields. In: Proceedings of the symposium on multivariate analysis
33. Bickel PJ (1974) Edgeworth expansions in nonparametric statistics. Ann Stat 2:1–20
34. Bickel PJ, Lehmann EL (1974) Measures of location and scale. In: Hajek J (ed) The proceedings of the Prague symposium on asymptotic statistics, vol 1. Charles University, Prague, pp 25–36

35. Bickel PJ (1975) One-step Huber estimates in the linear model. J Am Stat Assoc 70:428–434
36. Bickel PJ, Hammel E, O'Connell J (1975) Sex bias in graduate admissions: data from Berkeley. Science 187:398–404
37. Bickel PJ, Lehmann EL (1975) Descriptive statistics for nonparametric models. I. Introduction. Ann Stat 3:1038–1044
38. Bickel PJ, Lehmann EL (1975) Descriptive statistics for nonparametric models. II. Location. Ann Stat 3:1045–1069
39. Albers W, Bickel PJ, van Zwet WR (1976) Asymptotic expansions for the power of distribution-free tests in the one-sample problem. Ann Stat 4:108–156
40. Bickel PJ, Doksum K (1976) Mathematical statistics: basic ideas and selected topics. Holden-Day, San Francisco
41. Bickel PJ, Lehmann EL (1976) Descriptive statistics for nonparametric models. III. Dispersion. Ann Stat 4:1139–1158
42. Bickel PJ (1976) Another look at robustness: a review of reviews and some new developments. Scand J Stat 3:145–168
43. Bickel PJ, Yahav J (1977) On selecting a set of good populations. Statistical decision theory and related topics, II. Academic, New York, pp 37–55
44. Bickel PJ (1978) Using residuals robustly I: testing for heteroscedastisity, nonlinearity, nonadditivity. Ann Stat 6:266–291
45. Bickel PJ, van Zwet WR (1978) Asymptotic expansions for the power of distribution free tests in the two-sample problems. Ann Stat 6:937–1004
46. Bickel PJ, Herzberg A (1979) Robustness of design against autocorrelation in time I: asymptotic theory, optimality for location and linear regression. Ann Stat 7:77–95
47. Bickel PJ, Lehmann EL (1979) Descriptive statistics for non-parametric models. IV. Spread. Contributions to statistics. J. Hajek memorial volume. Academia Prague, pp 33–40
48. Bickel PJ, Freedman DA (1980) On Edgeworth expansions for the bootstrap. Technical report
49. Bickel PJ, van Zwet WR (1980) On a theorem of Hoeffding. Asymptotic theory of statistical tests and estimation (Hoeffding Festschrift). Academic, New York
50. Bickel PJ, Doksum KA (1981) An analysis of transformations revisited. J Am Stat Assoc 76:296–311
51. Bickel PJ (1981) Minimax estimation of the mean of a normal distribution when the parameter space is restricted. Ann Stat 9:1301–1309
52. Bickel PJ, Chibisov DM, van Zwet WR (1981) On efficiency of first and second order. Int Stat Rev 49:169–175
53. Bickel PJ (1981) Quelques aspects de la statistique robuste. Ecole d'Ete de probabilites de St. Flour. Springer-Verlag Lecture notes in mathematics
54. Bickel PJ, Lehmann EL (1981) A minimax property of the sample mean in finite populations. Ann Stat 9:1119–1122

55. Bickel PJ, Herzberg A, Schilling MF (1981) Robustness of design against autocorrelation in time II: optimality, theoretical and numerical results for the first-order autoregressive process. J Am Stat Assoc 76:870–877
56. Bickel PJ, Freedman DA (1981) Some asymptotic theory for the bootstrap. Ann Stat 9:1218–1228
57. Bickel PJ, Yahav JA (1982) Asymptotic theory of selection and optimality of Gupta's rules. In: Kallianpur G, Krishnaiah PR, Ghosh JK (eds) Statistics and probability: essays in Honor of C.R. Rao. North-Holland, pp 109–124
58. Bickel PJ, Robinson J (1982) Edgeworth expansions and smoothness. Ann Probab 10:500–503
59. Bickel PJ (1982) On adaptive estimation. Ann Stat 10:647–671
60. Bickel PJ (1982) Shifting integer valued random variables. A Festschrift for E.L. Lehmann, 49–61
61. Bickel PJ, Freedman D (1982) Bootstrapping regression models with many parameters. A. Festschrift for E.L. Lehmann, 28–48
62. Bickel PJ (1983) Minimax estimation of the mean of normal distribution subject to doing well at a point. Recent advances in statistics, a Festschrift for H. Chernoff. Academic
63. Bickel PJ, Breiman L (1983) Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. Ann Probab 11:185–214
64. Bickel PJ, Collins J (1983) Minimizing fisher information over mixtures of distributions. Sankhyā 45:1–19
65. Bickel PJ, Freedman D (1984) Asymptotic normality and the bootstrap in stratified sampling. Ann Stat 12:470–482
66. Bickel PJ (1984) Review: contributions to a general asymptotic statistical theory by J. Pfanzagl. Ann Stat 12:786–791
67. Bickel PJ (1984) Review: theory of point estimation by E.L. Lehmann. Metrika 256
68. Bickel PJ (1984) Robust regression based on infinitesimal neighbourhoods. Ann Stat 12:1349–1368
69. Bickel PJ (1984) Parametric robustness. Ann Stat 12:864–879
70. Bickel PJ, Yahav JA (1985) On estimating the number of unseen species: how many executions were there? Technical report
71. Bickel PJ, Götze F, van Zwet WR (1985) A simple analysis of third-order efficiency of estimates. In: Le Cam L, Olshen R (eds) Proceedings of the Berkeley conference in Honor of Jerzy Neyman and Jack Kiefer, vol. II. Wadsworth, pp 749–767
72. Bickel PJ, Klaassen CAJ (1986) Empirical Bayes estimation in functional and structural models and uniformly adaptive estimation of location. Adv Appl Math 7:55–69
73. Bickel PJ (1986) Efficient testing in a class of transformation models. In: Gill R, Voors MN (eds) Papers on semiparametric models at the ISI centenary session, Amsterdam. C.W.I. Amsterdam

74. Bickel PJ, Ritov J (1987) Efficient estimation in the errors in variables model. Ann Stat 15:513–541
75. Bickel PJ, Götze F, van Zwet WR (1987) The Edgeworth expansion for *U* statistics of degree 2. Ann Stat 15:1463–1484
76. Bickel PJ, Yahav JA (1987) On estimating the total probability of the unobserved outcomes of an experiment. In: van Ryzin J (ed) Adaptive statistical procedures and related topics. Institute of Mathematical Statistics, Hayward
77. Bickel PJ (1987) Robust estimation. In: Johnson N, Kotz S (eds) Encyclopedia of statistical sciences. Wiley, New York
78. Bickel PJ, Yahav JA (1988) Richardson extrapolation and the bootstrap. J Am Stat Assoc 83:387–393
79. Bickel PJ, Mallows C (1988) A note on unbiased Bayes estimates. Am Stat 42:132–134
80. Bickel PJ (1988) Estimating the size of a population. In: Proceedings of the IVth Purdue symposium on decision theory and related topics
81. Bickel PJ, Ritov Y (1988) Estimating integrated squared density derivatives. Sankhyā 50:381–393
82. Bickel PJ, Krieger A (1989) Confidence bands for a distribution function using the bootstrap. J Am Stat Assoc 84:95–100
83. Bai C, Bickel PJ, Olshen R (1989) The bootstrap for prediction. In: Proceedings of an oberwolfach conference. Springer
84. Bickel PJ, Ghosh JK (1990) A decomposition for the likelihood ratio statistic and the Bartlett correction—a Bayesian argument. Ann Stat 18:1070–1090
85. Bickel PJ, Ritov Y (1990) Achieving information bounds in non and semiparametric models. Ann Stat 18:925–938
86. Bickel PJ, Ritov Y (1991) Large sample theory of estimation in biased sampling regression models. I. Ann Stat 19:797–816
87. Bickel PJ, Ritov Y, Wellner J (1991) Efficient estimation of linear functionals of a probability measure *P* with known marginals. Ann Stat 19:1316–1346
88. Bickel PJ, Ritov Y (1992) Testing for goodness of fit: a new approach. In: Saleh AK Md E (ed) Nonparametric statistics and related topics. North Holland, Amsterdam, pp 51–57
89. Bickel PJ, Zhang P (1992) Variable selection in non-parametric regression with categorical covariates. J Am Stat Assoc 87:90–98
90. Bickel PJ (1992) Inference and auditing: the stringer bound. Int Stat Rev 60:197–209
91. Bickel PJ (1992) Theoretical comparison of bootstrap *t* confidence bounds. In: Billard L, Le Page R (eds) Exploring the limits of the bootstrap. Wiley, New York, pp 65–76
92. Bickel PJ, Millar PW (1992) Uniform convergence of probability measures on classes of functions. Stat Sin 2:1–15
93. Bickel PJ, Nair VN, Wang PCC (1992) Nonparametric inference under biased sampling from a finite population. Ann Stat 20:853–878

94. Bickel PJ (1993) Estimation in semiparametric models. In: Rao CR (ed) Chapter 3 in multivariate analysis: future directions, vol 5. pp 55–73. Elsevier Science Publishers, Amsterdam
95. Bickel PJ, Krieger A (1993) Extensions of Chebychev's inequality with applications. Probab Math Stat 13:293–310
96. Bickel PJ, Ritov Y (1993) Efficient estimation using both direct and indirect observations. Theory Probab Appl 38:194–213
97. Bickel PJ, Ritov Y (1993) Ibragimov Hasminskii models. In: Fifth Purdue international symposium on decision theory and related topics
98. Bickel PJ, Chernoff H (1993) Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In: Mitra SK (ed) Bahadur volume. Wiley Eastern, New Dehli
99. Bickel PJ, Ritov Y (1993) Discussion of papers by Feigelson and Nousek. In: Feigelson E, Babu GJ (eds) Statistical challenges in modern astronomy. Springer, New York
100. Bickel PJ, Klaassen C, Ritov Y, Wellner J (1993) Efficient and adaptive estimation for semiparametric models. Johns Hopkins University Press. Reprinted in 1998, Springer, New York
101. Bickel PJ, Ritov Y (1995) Estimating linear functionals of a PET image. IEEE Trans Med Imaging 14:81–88
102. Bickel PJ, Hengartner N, Talbot L, Shepherd I (1995) Estimating the probability density of the scattering cross section from Rayleigh scattering experiments. J Opt Soc Am A 12:1316–1323
103. Bickel PJ, Nair VN (1995) Asymptotic theory of linear statistics in sampling proportional to size without replacement. Probab Math Stat 15:85–99
104. Bickel PJ, Ritov Y (1995) An exponential inequality for U-statistics with applications to testing. Probab Eng Inf Sci 9:39–52
105. Bickel PJ, Fan J (1996) Some problems on estimation of unimodal densities. Stat Sin 6:23–45
106. Bickel PJ, Ren JJ (1996) The m out of n bootstrap and goodness of fit tests with doubly censored data. In: Rieder H (ed) Festschrift for P.J. Huber. Springer
107. Bickel PJ, Cosman PC, Olshen RA, Spector PC, Rodrigo AG, Mullins JI (1996) Covariability of V3 loop amino acids. AIDS Res Hum Retrovir 12:1401–1410
108. Bickel PJ, Bühlmann P (1996) What is a linear process? Proc Natl Acad Sci 93:12128–12131
109. Bickel PJ, Ritov Y (1996) Inference in hidden Markov models I: local asymptotic normality in the stationary case. Bernoulli 2:199–228
110. Bickel PJ, Götze F, van Zwet WR (1997) Resampling fewer than n observations: gains, losses, and remedies for losses. Stat Sin 1:1–31
111. Bickel PJ (1997) Discussion of statistical aspects of hipparcos photometric data (F. van Leeuwen et al.). In: Babu GJ, Fergelson Fr (eds) Statistical challenges in modern astronomy II. Springer
112. Bickel PJ (1997) An overview of SCMA II. In: Babu GJ, Fergelson Fr (eds) Statistical challenges in modern astronomy II. Springer

113. Bickel PJ (1997) LAN for ranks and covariates. In: Pollard D, Torgersen E, Yang G (eds) Festschrift for Lucien Le Cam. Springer
114. Bickel PJ, Bühlmann P (1997) Closure for linear processes. J Theor Probab 10:445–479
115. Bickel PJ, Petty KF, Jiang J, Ostland M, Rice J, Ritov Y, Schoenberg R (1998) Accurate estimation of travel times from single loop detectors. Transp Res A 32:1–18
116. Nielsen JP, Linton O, Bickel PJ (1998) On a semiparametric survival model with flexible covariate effect. Ann Stat 26:215–241
117. van der Laan MJ, Bickel PJ, Jewell NP (1998) Singly and doubly censored current status data: estimation, asymptotics and regression. Scand J Stat 24:289–307
118. Bickel PJ, Ritov Y, Rydén T (1998) Asymptotic normality of the maximum-likelihood estimator for general Hidden Markov models. Ann Stat 26:1614–1635
119. Bickel PJ, Bühlmann P (1999) A new mixing notion and functional central limit theorem for a sieve bootstrap in time series. Bernoulli 5:413–446
120. Sakov A, Bickel PJ (2000) An Edgeworth expansion for the m out of n bootstrapped median. Stat Probab Lett 49:217–223
121. Kwon J, Coifman B, Bickel PJ (2000) Day-to-day travel-time trends and travel-time prediction from loop-detector data. Transp Res Board Rec 1717: 120–129
122. Bickel PJ, Ritov Y (2000) Non- and semiparametric statistics: compared and contrasted. J Stat Plan Inference 91:209–228
123. Bickel PJ, Levina E (2001) The earth mover's distance is the Mallows distance: some insights from statistics. In: Proceedings of ICCV '01, Vancouver, pp 251–256
124. Aït-Sahalia Y, Bickel PJ, Stoker TM (2001) Goodness-of-fit tests for kernel regression with an application to option implied volatilities. J Econ 105:363–412
125. Bickel PJ, Ren JJ (2001) The bootstrap in hypothesis testing. In: State of the art in probability and statistics. IMS lecture notes-monograph series, vol 36, pp 91–112. Festschrift for W.R. van Zwet
126. Bickel PJ, Buyske S, Chang H, Ying Z (2001) On maximizing item information and matching difficulty with ability. Psychometrika 66:69–77
127. Bickel PJ, Kwon J (2001) Inference for semiparametric models: some questions and an answer (with discussion). Stat Sin 11:863–960
128. Bickel PJ, Doksum KA (2001) Mathematical statistics, vol 1: basic ideas and selected topics, 2nd edn. Prentice Hall, New Jersey
129. Bickel PJ, Lehmann EL (2001) Frequentist interpretation of probability. In: International encyclopedia of the social and behavioral sciences. Elsevier Science Ltd., Oxford, pp 5796–5798
130. Bickel PJ, Chen C, Kwon J, Rice J, Varaiya P, van Zwet E (2002) Traffic flow on a freeway network. Nonlinear estimation and classification, vol 171. Springer, pp 63–81

131. Bickel PJ, Sakov A (2002) Equality of types for the distribution of the maximum for two values of *n* implies extreme value type. Extremes 5:45–53
132. Petty K, Ostland M, Kwon J, Rice J, Bickel PJ (2002) A new methodology for evaluating incident detection algorithms. Transp Res C 10:189–204
133. Kwon J, Min K, Bickel PJ, Renne PR (2002) Statistical methods for jointly estimating the decay constant of ^{40}K and the age of a dating standard. Math Geol 34:457–474
134. Bickel PJ, Sakov A (2002) Extrapolation and the bootstrap. Sankhyā 64:640–652
135. Bickel PJ, Kechris KJ, Spector PC, Wedemayer GJ, Glazer AN (2002) Finding important sites in protein sequences. PNAS 99:14764–14771
136. Bickel PJ, Ritov Y, Rydén T (2002) Hidden Markov model likelihoods and their derivatives behave like i.i.d. ones. Annales de l'Institut Henri Poincare (B) Probability and Statistics 38:825–846
137. Berk R, Bickel PJ, Campbell K, Fovell R, Keller-McNulty S, Kelly E, Linn R, Park B, Perelson A, Rouphail N, Sacks J, Schoenberg F (2002) Workshop on statistical approaches for the evaluation of complex computer models. Stat Sci 17:173–192
138. Bickel PJ, Ritov Y (2003) Inference in hidden Markov models. In: Proceedings of the international congress of mathematicians, vol 2. World Publishers, Hong Kong
139. Kim N, Bickel PJ (2003) The limit distribution of a test statistic for bivariate normality. Stat Sin 13:327–349
140. Bickel PJ, Ritov Y (2003) Nonparametric estimators which can be "plugged-in". Ann Stat 31:1033–1053
141. Kechris KJ, van Zwet E, Bickel PJ, Eisen MB (2004) Detecting DNA regulatory motifs by incorporating positional trends in information content. Genome Biol 5:1–21
142. Ge Z, Bickel PJ, Rice JA (2004) An approximate likelihood approach to nonlinear mixed effects models via spline approximation. Comput Stat Data Anal 46:747–776
143. Bickel PJ (2004) Unorthodox bootstraps. J Korean Stat Soc 32:213–224
144. Chen A, Bickel PJ (2004) Robustness of prewhitening of heavy tailed sources. Springer Lect Notes Comput Sci
145. Bickel PJ, Levina E (2004) Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. Bernoulli 10:989–1010
146. Chen A, Bickel PJ (2005) Consistent independent component analysis and prewhitening. IEEE Trans Signal Process 53:3625–3633
147. Levina E, Bickel PJ (2005) Maximum likelihood estimation of intrinsic dimension. In: Saul LK, Weiss Y, Bottou L (eds) Advances in NIPS, vol 17
148. Olshen AB, Cosman PC, Rodrigo AG, Bickel PJ, Olshen RA (2005) Vector quantization of amino acids: analysis of the HIV V3 loop region. J Stat Plan Inference 130:277–298

149. van Zwet EW, Kechris KJ, Bickel PJ, Eisen MB (2005) Estimating motifs under order restrictions. Stat Appl Genet Mol Biol 4:1–16. Art.1
150. Bickel PJ, Ritov Y, Zakai A (2006) Some theory for generalized boosting algorithms. JMLR 7:705–732
151. Kechris KJ, Lin JC, Bickel PJ, Glazer AN (2006) Quantitative exploration of the occurrence of lateral gene transfer by using nitrogen fixation genes as a case study. PNAS 103:9584–9589
152. Bickel PJ, Li B (2006) Regularization in statistics (with discussion). Test 15:271–344
153. Levina E, Bickel PJ (2006) Texture synthesis and nonparametric resampling of random fields. Ann Stat 34:1751–1773
154. Bickel PJ, Ritov Y, Stoker TM (2006) Tailor-made tests for goodness-of-fit to semiparametric hypotheses. Ann Stat 34:721–741
155. Chen A, Bickel PJ (2006) Efficient independent component analysis. Ann Stat 34:2825–2855
156. Bickel PJ, Li B (2007) Local polynomial regression on unknown manifolds. In: Complex datasets and inverse problems: tomography, networks and beyond. IMS lecture notes-monograph series, vol 54, pp 177–186
157. The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816
158. Margulies EH et al (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res 17:760–774
159. Bickel PJ, Kleijn B, Rice J (2007) On detecting periodicity in astronomical point processes. In: Challenges in modern astronomy IV: ASP conference series, vol 371
160. Bickel PJ, Chen C, Kwon J, Rice J, van Zwet E, Varaiya P (2007) Measuring traffic. Stat Sci 22:581–597
161. Bickel PJ, Levina E (2008) Regularized estimation of large covariance matrices. Ann Stat 36:199–227
162. Bengtsson T, Bickel PJ, Li B (2008) Curse-of-dimensionality revisited: collapse of the particle filter in very large scale systems. In: IMS collections: probability and statistics: essays in Honor of David A. Freedman, vol 2, pp 316–334
163. Bickel PJ, Li B, Bengtsson T (2008) Sharp failure rates for the bootstrap particle filter in high dimensions. In: IMS collections: pushing the limits of contemporary statistics: contributions in Honor of Jayanta K. Ghosh, vol 3, pp 318–329
164. Bickel PJ, Sakov A (2008) On the choice of m in the m out of n bootstrap and its application to confidence bounds for extreme percentiles. Stat Sin 18:967–985
165. Rothman AJ, Bickel PJ, Levina E, Zhu J (2008) Sparse permutation invariant covariance estimation. Electron J Stat 2:494–515

166. Snyder C, Bengtsson T, Bickel PJ, Anderson J (2008) Obstacles to high-dimensional particle filtering. Mon Weather Rev 136:4629–4640
167. Bickel PJ, Kleijn B, Rice J (2008) Event-weighted tests for detecting periodicity in photon arrival times. Astrophys J 685:384–389
168. Bickel PJ, Levina E (2008) Covariance regularization by thresholding. Ann Stat 36:2577–2604
169. Bickel PJ, Yan D (2008) Sparsity and the possibility of inference. Sankhyā 70:1–24
170. Bickel PJ, Ritov Y, Tsybakov AB (2009) Simultaneous analysis of Lasso and Dantzig selector. Ann Stat 37:1705–1732
171. Meinshausen N, Bickel PJ, Rice J (2009) Efficient blind search: optimal power of detection under computational cost constraints. Ann Appl Stat 3:38–60
172. Bickel PJ, Brown JB, Huang H, Li Q (2009) An overview of recent developments in genomics and associated statistical methods. Philos Trans R Soc A 367:4313–4337
173. MacArthur S et al (2009) Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions Genome Biol 10(7), Art. R80
174. Bickel PJ, Chen A (2009) A nonparametric view of network models and Newman-Girvan and other modularities. PNAS 106:21068–21073
175. Bahadur RR, Bickel PJ (2009) An optimality property of Bayes' test statistics. In: Optimality: the third Erich L. Lehmann symposium. IMS lecture notes-monograph series, vol 57, pp 18–30
176. Bickel PJ, Ritov Y, Tsybakov AB (2010) Hierarchical selection of variables in sparse high-dimensional regression. In: IMS collections: borrowing strength: theory powering applications – a Festschrift for Lawrence D. Brown, vol. 6, pp 56–69
177. Lei J, Bickel PJ, Snyder C (2010) Comparison of ensemble Kalman filters under non-Gaussianity. Mon Weather Rev 138:1293–1306
178. Bickel PJ, Boley N, Brown JB, Huang H, Zhang NR (2010) Subsampling methods for genomic inference. Ann Appl Stat 4:1660–1697
179. Xu N, Bickel PJ, Huang H (2010) Genome-wide detection of transcribed regions through multiple RNA tiling array analysis. Int J Syst Synth Biol 1:155–170
180. Aswani A, Bickel PJ, Tomlin, C (2011) Regression on manifolds: estimation of the exterior derivative. Ann Stat 39:48–81
181. Li Q, Brown JB, Huang H, Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. Ann Appl Stat 5:1752–1779
182. Li JJ, Jiang CR, Brown JB, Huang H, Bickel PJ (2011) Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. PNAS 108:19867–19872
183. Graveley BR et al (2011) The developmental transcriptome of *Drosophila melanogaster*. Nature 471:473–479

184. Bickel PJ, Gel YR (2011) Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series. J R Stat Soc B 73:711–728
185. Hoskins RA et al (2011) Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. Genome Res 21:182–192

- **Comments, Discussions and Other Publications**

1. Bickel PJ (1979) Comment on "Conditional independence in statistical theory" by Dawid, A.P. J R Stat Soc B 41:21
2. Bickel PJ (1979) Comment on "Edgeworth and saddle point approximations with statistical applications" by Barndorff-Nielsen, O. and Cox, D.R. J R Stat Soc B 41:307
3. Bickel PJ (1980) Comment on "Sampling and Bayes' inference in scientific modelling and robustness" by Box, G. J R Stat Soc A 143:383–431
4. Bickel PJ (1983) Comment on "Bounded influence regression" by Huber, P.J. J Am Stat Assoc 78:75–77
5. Bickel PJ (1984) Comment on "Analysis of transformed data" by Hinkley, D. and Runger, S. J Am Stat Assoc 79:309
6. Bickel PJ (1984) Comment on "Adaptive estimation of nonlinear regression models" by Manski, C.F. Econ Rev 3:145–210
7. Bickel PJ (1987) Comment on "Better bootstrap confidence intervals" by Efron, B. J Am Stat Assoc 82:191
8. Bickel PJ (1988) Mathematical sciences: some research trends (section on statistics). National Academy Press, Washington
9. Bickel PJ (1988) Comment on "Theoretical comparison of bootstrap confidence intervals" by Hall, P. Ann Stat 16:959–961
10. Bickel PJ (1988) Comment on "Rank based robust analysis of models" by Draper, D. Stat Sci
11. Bickel PJ, Ritov Y (1990) Comment on "A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography" by Silverman, B.W., Jones, M.C., Wilson, J.D. and Nychka, D.W. J R Stat Soc B 52:271–325 (with discussion)
12. Bickel PJ, Le Cam L (1990) A conversation with Ildar Ibragimov. Stat Sci 5:347–355
13. Bickel PJ (1990) Renewing US mathematics: a plan for the 1990s. Report of NRC Committee. N.A.S. Press
14. Bickel PJ (1991) Comment on "Fisherian inference in likelihood and prequential frames of reference" by Dawid, A.P. J R Stat Soc B 53:105
15. Bickel PJ (1994) What academia needs. Am Stat 49:1–6
16. Bickel PJ, Ostland M, Petty K, Jiang J, Rice J, Schoenberg R (1997) Simple travel time estimation from single trap loop detectors. Intellimotion 6:4–5

17. Bickel PJ (1997) Discussion of "The evaluation of forensic DNA evidence". PNAS 94:5497
18. Bickel PJ (2000) Comment on "Hybrid resampling methods for confidence intervals" by Chuang, C.S. and Lai, T.L. Stat Sin 10:37–38
19. Bickel PJ, Ritov Y (2000) Comment on "On profile likelihood" by Murphy, S.A. and van der Vaart, A.W. J Am Stat Assoc 95:466–468
20. Bickel PJ (2000) Statistics as the information science. In: Opportunities for the mathematical sciences, NSF workshop, pp 9–11
21. Bickel PJ, Lehmann EL (2001) Frequentist inference. In: International encyclopedia of the social and behavioral sciences. Elsevier Science Ltd., Oxford, pp 5789–5796
22. Bickel PJ (2002) Comment on "What is a statistical model?" by McCullagh, P. Ann Stat 30:1225–1310
23. Bickel PJ (2002) Comment on "Box-Cox transformations in linear models: large sample theory and tests of normality" by Chen, G., Lockhart, R.A. and Stephens, M.A. Can J Stat 30:177–234
24. Bickel PJ (2002) The board on mathematical sciences has evolved. IMS Bull
25. Bickel PJ, Ritov Y (2004) The golden chain: discussion of three boosting papers. Ann Stat 32:91–96
26. Bickel PJ (2005) Mathematics and 21st century biology. NRC
27. Bickel PJ (2007) Comment on "Maximum likelihood estimation in semiparametric regression models with censored data" by Zeng, D. and Lin, D.Y. J R Stat Soc B 69:546–547
28. Bickel PJ, Ritov Y (2008) Discussion of "Treelets—an adaptive multi-scale basis for sparse unordered data" by Lee, A.B., Nadler, B. and Wasserman, L. Ann Appl Stat 2:474–477
29. Bickel PJ, Ritov Y (2008) Discussion of Mease and Wyner: and yet it overfits. JMLR 9:181–186
30. Bickel PJ, Xu Y (2009) Discussion of: Brownian distance covariance. Ann Appl Stat 3:1266–1269
31. Bickel PJ (2010) Review of Huber: robust statistics. SIAM Rev

- **Working Papers**

1. Lei J, Bickel PJ (2009) Ensemble filtering for high dimensional nonlinear state space models. Mon Weather Rev, to appear
2. Atherton J et al (2010) A model for Sequential Evolution of Ligands by EXponential enrichment (SELEX) data. Manuscripts
3. Bickel PJ, Lindner M (2010) Approximating the inverse of banded matrices by banded matrices with applications to probability and statistics. Theory Probab Appl, to appear
4. Kleijn BJK, Bickel PJ (2010) The semiparametric Bernstein-Von Mises theorem. Ann Stat, to appear

5. Song S, Bickel PJ (2011) Large vector auto regressions. Manuscripts
6. Bickel PJ, Chen A, Levina E (2011) The method of moments and degree distributions for network models. Ann Stat, to appear

Ph.D. Students of Peter J. Bickel

1. 1966: Dattaprabhakar Gokhale, “Some Problems in Independence and Dependence.”
2. 1967: Hira Lal Koul, “Estimation by Method of Ranks in Regression Models.”
3. 1968: Jan Geertsema, “Sequential Confidence Intervals Based on Rank Tests.”
4. 1969: Radhakrishnan Aiyar, “On Some Tests for Trend and Autocorrelation.”
5. 1970: Luis Bernabe Boza, “Asymptotically Optimal Tests for Finite Markov Chains.”
6. 1971: Barry Rees James, “A Functional Law of the Iterated Logarithm for Weighted Empirical Distributions.”
7. 1971: Djalma Pessoa, “Asymptotically Minimax Fixed Length Confidence Intervals.”
8. 1972: Eduardo De Weerth, “Sequential Estimation of a Truncation Parameter.”
9. 1973: John Collins, “Robust Estimation of a Location Parameter in the Presence of Asymmetry.”
10. 1973: Jose Dachs, “Asymptotic Expansions for M-Estimators.”
11. 1974: Steinar Bjerve, “Error Bounds and Asymptotic Expansions for Linear Combinations of Order-Statistics.”
12. 1974: Olivier Muron, “Asymptotic Approximations of the Characteristics of Sequential Bounded Length Confidence Intervals.”
13. 1974: Jeffrey Polovina, “The Estimation of Simple Linear Regression Coefficients from Incomplete Data.”
14. 1974: Pham Xuan Quang, “Robust Sequential Testing.”
15. 1976: Winston Chow, “A New Method of Approximation to Various Distributions Arising in Testing Problems.”
16. 1976: Aldo Viollaz, “Nonparametric Estimation of Probability Density Functions Using Orthogonal Expansions.”
17. 1976: Chien-Fu Wu, “Contributions to Optimization Theory with Applications to Optimal Design of Experiments.”

18. 1977: Jeyaraj Vadiveloo, "On the Theory of Modified Randomization Tests for Nonparametric Hypotheses."
19. 1978: Joel Brodsky, "On Estimating a Common Mean."
20. 1978: Thomas Hammerstrom, "On Asymptotic Optimality Properties of Tests and Estimates in the Presence of Increasing Numbers of Nuisance Parameters."
21. 1978: Nilson Marcondes, "Estimation of Multivariate Densities, Conditional Densities and Related Functions."
22. 1979: Eugene Poggio, "Accuracy Functions for Confidence Bounds: A Basis for Sample Size Determination."
23. 1979: Mark Schilling, "Testing for Goodness of Fit Based on Nearest Neighbors."
24. 1979: Paul Wang, "Asymptotic Robust Tests in the Presence of Nuisance Parameters."
25. 1980: Ronaldo Iachan, "Topics on Systematic Sampling."
26. 1981: Ian Abramson, "On Kernel Estimates of Probability Densities."
27. 1981: Robert Holmes, Jr., "Contributions to the Theory of Parametric Estimation in Randomly Censored Data."
28. 1982: Donald Andrews, "A Model for Robustness against Distributional Shape and Dependence over Time."
29. 1983: Michael Trosset, "Minimax Estimation with Side Conditions."
30. 1983: Ya'acov Ritov, "Quasi Bayesian Robust Inference." (at the Hebrew University of Jerusalem, co-advised by J. Yahav.)
31. 1984: Enio Jelihovschi, "Estimation of Poisson Parameters Subject to Constraints."
32. 1987: Julian Faraway, "Smoothing in Adaptive Estimation."
33. 1987: Byeong Uk Park, "Efficient Estimation in the Two-Sample Semiparametric Location Scale Model and the Orientation Shift Model."
34. 1989: Jianqing Fan, "Contributions to the Estimation of Nonregular Functionals."
35. 1989: Moxiu Mo, "Robust Additive Regression."
36. 1990: Kun Jin, "Empirical Smoothing Parameter Selection in Adaptive Estimation."
37. 1990: Yonghua Wang, "On Efficient Estimation Under Equation Constraints."
38. 1990: Ping Zhang, "Variable Selection in Nonparametric Regression."
39. 1991: Alex Bajamonde, "On Efficient and Robust Estimation in Semiparametric Linear Regression Models with Missing Data."
40. 1991: Panagiotis Lorentziadis, "Forecasts in Oil Exploration and Prospect Evaluation for Financial Decisions: A Semiparametric Approach."
41. 1992: Zaiqian Shen, "Robust Estimation in Semiparametric Models."
42. 1992: Yazhen Wang, "Nonparametric Estimation Subject to Shape Restrictions."
43. 1993: Niklaus Hengartner, "Topics in Density Estimation."

44. 1993: Mark van der Laan, “Efficient and Inefficient Estimation in Semiparametric Models.” (at the University of Utrecht, co-advised by Richard Gill.)
45. 1994: Namhyun Kim, “Goodness of Fit Test in Multivariate Normal Distributions.”
46. 1995: Jiming Jiang, “REML estimation: Asymptotic behavior and related topics.”
47. 1998: Zhiyu Ge, “The Histogram Method and the Conditional Maximum Profile Likelihood Method for Nonlinear Mixed Effects Models.”
48. 1998: Anat Sakov, “Using the m out of n Bootstrap in Hypothesis Testing.”
49. 2000: Yoram Gat, “Overfit Bounds for Classification Algorithms.”
50. 2000: Jaimyoung Kwon, “Calculus of Statistical Efficiency in a General Setting; Kernel Plug-in Estimation for Markov Chains; Hidden Markov Modeling of Freeway Traffic.”
51. 2002: Jenher Jeng, “Wavelet Methodology for Advanced Nonparametric Curve Estimation: from Confidence Band to Sharp Adaptation.”
52. 2002: Elizaveta Levina, “Statistical Issues in Texture Analysis.”
53. 2003: Katherina Kechris, “Statistical Methods for Discovering Features in Molecular Sequences.”
54. 2004: Aiyou Chen, “Semiparametric Inference for Independent Component Analysis.”
55. 2006: Bo Li, “On Goodness-of-fit Tests of Semiparametric Models.”
56. 2008: Choongsoon Bae, “Analyzing Random Forests.”
57. 2008: Na Xu, “Transcriptome Detection by Multiple RNA Tiling Array Analysis and Identifying Functional Conserved Non-coding Elements by Statistical Testing.”
58. 2008: Donghui Yan, “Some Issues with Dimensionality in Statistical Inference.”
59. 2009: James Brown, “Mapping the Affinities of Sequence-specific DNA-binding Proteins.”
60. 2010: Jing Lei, “Non-linear Filtering for State Space Models - High-dimensional Applications and Theoretical Results.”
61. 2011: Ying Xu, “Regularization Methods for Canonical Correlation Analysis, Matrices and Renyi Correlation.”

Photos of Peter J. Bickel

Madeleine, Eliezer and Peter Bickel 1941, Romania.

Toronto, Canada, 1951.

The Bickel family, 1971. From left to right Amanda, Nancy, Peter, and Steve.

Bryce Crawford, Home Secretary of the National Academy of Sciences, and Peter Bickel signing the book at the National Academy of Sciences ceremony April 29, 1987. Photograph by the National Academy of Sciences.

Madeleine Korb, Peter's mother, Nancy Bickel, and Peter Bickel at celebrations at National Academy of Sciences, April 1987.

Katerina Kechris, Haiyan Huang, Friedrich Goetze, Bickel, behind, Anton Shick, at the Symposium on Frontiers of Statistics in honor of Peter Bickel's 65th birthday at Princeton University, 2006.

David Donoho, Iain Johnstone and Peter Bickel at the National Academy of Sciences meeting, 2006.

Peter Bickel and his former students, colleagues, and friends at "Symposium on Frontiers of Statistics in honor of Peter Bickel's 65th birthday" at Princeton University, 2006.

Nancy Bickel and Peter Bickel at the "Symposium on Frontiers of Statistics in honor of Peter Bickel's 65th birthday" at Princeton University, 2006.

From Left: Her excellency Cora Minderhoud, Consul General of the Netherlands in New York, Willem van Zwet, Jianqing Fan with Peter Bickel after he was appointed Commander in the Order of Oranje-Nassau, May 19, 2006, Princeton University.

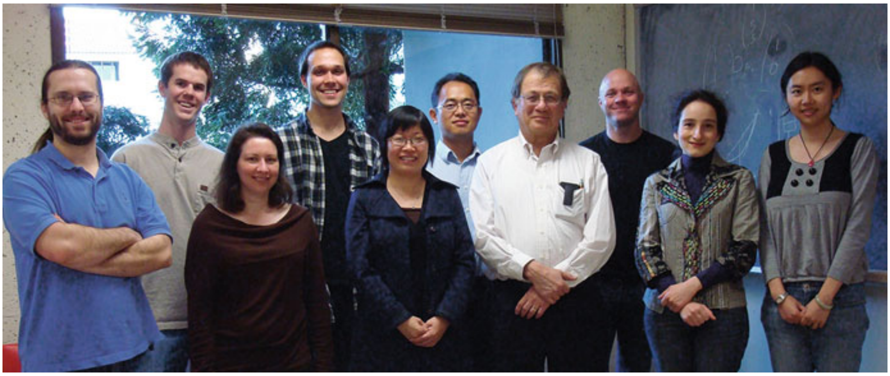
Peter and Nancy Bickel and his former student Liza Levina and her husband Edward Ionides at "Symposium on Frontiers of Statistics in honor of Peter Bickel's 65th birthday" at Princeton University, 2006.

Ya'acov Ritov, Ilana Ritov, and Peter Bickel, 2008, in the old city of Jerusalem.

Peter Bickel and Jianqing Fan, 2009, the Yellow River in Jinan, China.

Peter Bickel and Noureddine El Karoui, Spring 2012.

The Bickel family in Greece. Left to right back row Peter Mayer, Eliyana Adler, Peter Bickel. Front row, Amanda Bickel, Rana, Maya, and Selah Bickel, daughters of Eliyana Adler and Stephen Bickel, Nancy Bickel, Zachary Mayer-Bickel, Stephen Bickel holding Miles Mayer-Bickel, Photograph taken by Peter Mayer near Lindos, Rhodes, Greece, July, 2011.



From left: Ben Brown, Marcus Stoiber, Taly Arbel, Garrett Robinson, Haiyan Huang, Hao Xiong, Peter Bickel, Nathan Boley, Maya Polishchuk, and Jessica Li. Bickel's bioinformatics group, 2012.

Chapter 1
Rank-Based Nonparametrics

Willem R. van Zwet

1.1 Introduction to Two Papers on Higher Order Asymptotics

1.1.1 Introduction

Peter Bickel has contributed substantially to the study of rank-based nonparametric statistics. Of his many contributions to research in this area I shall discuss his work on second order asymptotics that yielded surprising results and set off more than a decade of research that deepened our understanding of asymptotic statistics. I shall restrict my discussion to two papers, which are Albers et al. (1976) "Asymptotic expansions for the power of distribution free tests in the one-sample problem" and Bickel (1974) "Edgeworth expansions in nonparametric statistics" where the entire area is reviewed.

1.1.2 Asymptotic Expansions for the Power of Distribution Free Tests in the One-Sample Problem

Let $X_1, X_2, \cdots$ be i.i.d. random variables with a common distribution function F_θ for some real-valued parameter θ. For $N = 1, 2, \cdots$, let A_N and B_N be two tests of level $\alpha \in (0,1)$ based on $X_1, X_2, \cdots, X_N$ for the null-hypothesis $H: \theta = 0$ against a contiguous sequence of alternatives $K_{N,c}: \theta = cN^{-1/2}$ for a fixed $c > 0$. Let $\pi_{A,N}(c)$ and $\pi_{B,N}(c)$ denote the powers of A_N and B_N for this testing problem and suppose

W.R. van Zwet (✉)
Leiden University, P.O.Box 9512, 2300 RA, Leiden, The Netherlands
e-mail: vanzwet@math.leidenuniv.nl

J. Fan et al. (eds.), *Selected Works of Peter J. Bickel*, Selected Works in Probability and Statistics 13, DOI 10.1007/978-1-4614-5544-8_1,

that A_N performs at least as well as B_N, i.e. $\pi_{A,N}(c) \geq \pi_{B,N}(c)$. Then we may look for a sample size $k = k_N \geq N$ such that B_k performs as well against alternative $K_{N,c}$ as A_N does for sample size N, i.e. $\pi_{B,k}(c(k/N)^{1/2}) = \pi_{A,N}(c)$. For finite sample size N it is generally impossible to find a usable expression for $k = k_N$, so one resorts to large sample theory and defines the asymptotic relative efficiency (ARE) of sequence $\{B_N\}$ with respect to $\{A_N\}$ as

$$e = e(B,A) = \lim_{N\to\infty} N/k_N.$$

If $\pi_{A,N}(c) \to \pi_A(c)$ and $\pi_{B,N}(c) \to \pi_B(c)$ uniformly for bounded c, and π_A and π_B are continuous, then e is the solution of

$$\pi_B(ce^{-1/2}) = \pi_A(c).$$

Since we assumed that A_N performs at least as well as B_N, we have $e \leq 1$.

If $e < 1$, the ARE provides a useful indication of the quality of the sequence $\{B_N\}$ as compared to $\{A_N\}$. To mimic the performance of A_N by B_k we need $k_N - N = N(1-e)/e + o(N)$ additional observations where the remainder term $o(N)$ is relatively unimportant. If $e = 1$, however, all we know is that the number of additional observations needed is $o(N)$, which may be of any order of magnitude, such as 1 or $N/\log\log N$. Hence in Hodges and Lehmann (1970) the authors considered the case $e = 1$ and proposed to investigate the asymptotic behavior of what they named the deficiency of B with respect to A

$$d_N = k_N - N,$$

rather than k_N/N. Of course this is a much harder problem than determining the ARE. To compute e, all we have to show is that $k_N = N/e + o(N)$, and only the limiting powers π_A and π_B enter into the solution. If $e = 1$, then $k_N = N + o(N)$, but for determining the deficiency, we need to evaluate k_N to the next lower order, which may well be $O(1)$ in which case we have to evaluate k_N with an error of the order $o(1)$. To do this, one will typically need asymptotic expansions for the power functions $\pi_{A,N}$ and $\pi_{B,N}$ with remainder term $o(N^{-1})$. For this we need similar expansions for the distribution functions of the test statistics of the two tests under the hypothesis as well as under the alternative.

In their paper Hodges and Lehmann computed deficiencies for some parametric tests and estimators, but they clearly had a more challenging problem in mind. When Frank Wilcoxon introduced his one- and two-sample rank tests (Wilcoxon 1945) most people thought that replacing the observations by ranks would lead to a considerable loss of power compared to the best parametric procedures. Since then, research had consistently shown that this is not the case. Many rank tests have ARE 1 when compared to the optimal test for a particular parametric problem, so it was not surprising that the first question that Hodges and Lehmann raised for further research was: “What is the deficiency (for contiguous normal shift alternatives) of the normal scores test or of van der Waerden’s X-test with respect to the t-test?”.

In the paper under discussion this question is generalized to other distributions than the normal and answered for the appropriate one-sample rank test as compared with the optimal parametric test. Let $X_1, X_2, \cdots, X_N$ be i.i.d. with a common distribution function G and density g, and let $Z_1 < Z_2 < \cdots < Z_N$ be the order statistics of the absolute values $|X_1|, |X_2|, \cdots, |X_N|$. If $Z_j = |X_{R(j)}|$, define $V_j = 1$ if $X_{R(j)} > 0$ and $V_j = 0$ otherwise. Let $a = (a_1, a_2, \cdots, a_N)$ be a vector of scores and define

$$T = \sum_{1 \le j \le N} a_j V_j. \tag{1.1}$$

T is the linear rank statistic for testing the hypothesis that g is symmetric about zero. Note that the dependence of G, g and a on N is suppressed in the notation. Conditionally on Z, the random variables $V_1, V_2, \cdots, V_N$ are independent with

$$P_j = P(V_j = 1|Z) = g(Z_j)/\{g(Z_j) + g(-Z_j)\}. \tag{1.2}$$

Under the null hypothesis, $V_1, V_2, \cdots, V_N$ are i.i.d. with $P(V_j = 1) = 1/2$. Hence the obvious strategy for obtaining an expansion for the distribution function of T is to introduce independent random variables $W_1, W_2, \cdots, W_N$ with $p_j = P(W_j = 1) = 1 - P(W_j = 0)$ and obtain an expansion for the distribution function of $\sum_{1 \le j \le N} a_j W_j$. In this expansion we substitute the random vector $P = (P_1, P_2, \cdots, P_N)$ for $p = (p_1, p_2, \cdots, p_N)$. The expected value of the resulting expression will then yield an expansion for the distribution function of T.

This approach is not without problems. Consider i.i.d. random variables $Y_1, Y_2, \cdots, Y_N$ with a common continuous distribution with mean $EY_j = 0$, variance $EY_j^2 = 1$, third and fourth moments $\mu_3 = EY_j^3$ and $\mu_4 = EY_j^4$, and third and fourth cumulants $\kappa_3 = \mu_3$ and $\kappa_4 = \mu_4 - 3\mu_2^2$. Let $S_N = N^{-1/2} \sum_{1 \le j \le N} Y_j$ denote the normalized sum of these variables. In Edgeworth (1905) the author provided a formal series expansion of the distribution function $F_N(x) = P(S_N \le x)$ in powers of $N^{-1/2}$. Up to and including the terms of order 1, $N^{-1/2}$ and N^{-1}, Edgeworth's expansion of $F_N(x)$ reads

$$\begin{aligned} F_N^*(x) = \Phi(x) - \phi(x) \cdot [&(\kappa_3/6)(x^2 - 1)N^{-1/2} \\ &+ \{(\kappa_4/24)(x^3 - 3x) + (\kappa_3^2/72)(x^5 - 10x^3 + 15x)\}N^{-1}]. \end{aligned} \tag{1.3}$$

We shall call this the three-term Edgeworth expansion. Though it was a purely formal series expansion, the Edgeworth expansion caught on and became a popular tool to approximate the distribution function of any sequence of continuous random variables U_N with expected value 0 and variance 1 that was asymptotically standard normal. As $\lambda_{3,N} = \kappa_3 N^{-1/2}$ and $\lambda_{4,N} = \kappa_4 N^{-1}$ are the third and fourth cumulants of the random variable S_N under discussion, one merely replaced these quantities by the cumulants of U_N in (1.3). Incidentally, I recently learned from Professor Ibragimov that the Edgeworth expansion was first proposed in Chebyshev (1890), which predates Edgeworth's paper by 15 years. Apparently this is one more example of Stigler's law of eponymy, which states that no scientific discovery – including Stigler's law – is named after its original discoverer (Stigler 1980).

A proof of the validity of the Edgeworth expansion for normalized sums S_N was given by Cramér (cf. 1937; Feller 1966). He showed that for the three-term Edgeworth expansion (1.3), the error $F_N^*(x) - F_N(x) = o(N^{-1})$ uniformly in x, provided that $\mu_4 < \infty$ and the characteristic function $\psi(t) = E\exp\{itY_j\}$ satisfies Cramér's condition

$$\limsup_{|t|\to\infty} |\psi(t)| < 1. \tag{1.4}$$

Assumption (1.4) can not be satisfied if Y_1 is a discrete random variable as then its characteristic function is almost periodic and the lim sup equals 1. In the case we are discussing, the summands a_jW_j of the statistic $\sum_{1\le j\le N} a_jW_j$ are independent discrete variables taking only two values 0 and a_j. However, the summands are not identically distributed unless the a_j as well as the p_j are equal. Hence the only case where the summands are i.i.d. is that of the sign test under the null-hypothesis, where $a_j = 1$ for all j, and the values 0 and 1 are assumed with probability $1/2$. In that case the statistic $\sum_{1\le j\le N} a_jW_j$ has a binomial distribution with point probabilities of the order $N^{-1/2}$ and it is obviously not possible to approximate a function F_N with jumps of order $N^{-1/2}$ by a continuous function F_N^* with error $o(N^{-1})$.

In all other cases the summands a_jW_j of $\sum_{1\le j\le N} a_jW_j$ are independent but not identically distributed. Cramér has also studied the validity of the Edgeworth expansion for the case that the Y_j are independent by not identically distributed. Assume again that $EY_j = 0$ and define S_N as the normalized sum $S_N = \sigma^{-1}\sum_{1\le j\le N} Y_j$ with $\sigma^2 = \sum_{1\le j\le N} EY_j^2$. As before $F_N(x) = P(S_N \le x)$ and in the three-term Edgeworth expansion $F_N^*(x)$ we replace $\kappa_3 N^{-1/2}$ and $\kappa_4 N^{-1}$ by the third and fourth cumulants of S_N. Cramér's conditions to ensure that $F_N^*(x) - F_N(x) = o(N^{-1})$ uniformly in x, are uniform versions of the earlier ones for the i.i.d. case: $EY_j^2 \ge c > 0, EY_j^4 \le C < \infty$ for $j = 1, 2, \cdots, N$, and for every $\delta > 0$ there exists $q_\delta < 1$ such that the characteristic functions $\psi_j(t) = E\exp\{itY_j\}$ satisfy

$$\sup_{|t|\ge\delta} |\psi_j(t)| < q_\delta \quad \text{for all } j. \tag{1.5}$$

As the a_jW_j are lattice variables (1.5) does not hold for even a single j and the plan of attack of this problem is beginning to look somewhat dubious. However, Feller points out, condition (1.5) is "extravagantly luxurious" for validating the three-term Edgeworth expansion and can obviously be replaced by $\sup_{|t|\ge\delta} |\Pi_{1\le j\le N}\psi_j(t)| = o(N^{-1})$ (cf. Feller 1966, Theorem XVI.7.2 and Problem XVI.8.12). This, in turn, is slightly too optimistic but it is true that the condition

$$\sup_{\delta\le|t|\le N} |\Pi_{1\le j\le N}\psi_j(t)| = o((N\log N)^{-1}) \tag{1.6}$$

is sufficient and the presence of $\log N$ is not going to make any difference. Hence (1.6) has to be proved for the case where $Y_j = a_j(W_j - p_j)$ and $S_N = \sum_{1\le j\le N} a_j(W_j - p_j)/\tau(p)$ with $\tau(p)^2 = \sum_{1\le j\le N} p_j(1-p_j)a_j^2$ and $\rho(t) = \prod_{1\le j\le N}\psi_j(t)$ is the characteristic function of S_N.

This problem is solved in Lemma 2.2 of the paper. The moment assumptions (2.15) of this lemma simply state that $N^{-1}\tau(p)^2 \ge c > 0$ and $N^{-1}\sum_{1\le j\le N} a_j^4 \le C < \infty$, and assumption (2.16) ensures the desired behavior of $|\prod_{1\le j\le N}\psi_j(t)|$ by requiring that there exist $\delta > 0$ and $0 < \varepsilon < 1/2$ such that

$$\lambda\{x : \exists j : |x - a_j| < \zeta,\ \varepsilon \le p_j \le 1-\varepsilon\} \ge \delta N\zeta \quad \text{for some } \zeta \ge N^{-3/2}\log N, \tag{1.7}$$

where λ is Lebesgue measure. This assumption ensures that the set of the scores a_j for which p_j is bounded away from 0 and 1, does not cluster too much about too few points. As is shown in the proof of Lemma 2.2 and Theorem 2.1 of the paper, assumptions (2.15) and (2.16) imply

$$\sup_{\delta\le|t|\le N} \Big|\prod_{1\le j\le N}\psi_j(t)\Big| \le \exp\{-d(\log N)^2\} = N^{-d\log N}, \tag{1.8}$$

which obviously implies (1.6). Hence the three-term Edgeworth expansion for $S_N = \sum_{1\le j\le N} a_j(W_j - p_j)/\tau(p)$ is valid with remainder $o(N^{-1})$, and in fact $O(N^{-5/4})$. This was a very real extension of the existing theory at the time.

To obtain an expansion for the distribution of the rank statistic $T = \sum_{1\le j\le N} a_j V_j$, the next step is to replace the probabilities p_j by the random quantities P_j in (1.2) and take the expectation. Under the null-hypothesis that the density g of the X_j is symmetric this is straightforward because $P_j = 1/2$ for all j. The alternatives discussed in the paper are contiguous location alternatives where $G(x) = F(x-\theta)$ for a specific known F with symmetric density f and $0 \le \theta \le CN^{-1/2}$ for a fixed $C > 0$. Finding an expansion for the distribution of T under these alternatives is highly technical and laborious, but fairly straightforward under the assumptions $N^{-1}\sum_{1\le j\le N} a_j^2 \ge c$, $N^{-1}\sum_{1\le j\le N} a_j^4 \le C$,

$$\lambda\{x : \exists j : |x - a_j| < \zeta\} \ge \delta N\zeta \quad \text{for some } \zeta \ge N^{-3/2}\log N \tag{1.9}$$

and some technical assumptions concerning f and its first four derivatives. Among many other things, the latter ensure that $\varepsilon \le P_j \le 1-\varepsilon$ for a substantial proportion of the P_j. Having obtained expansions for the distribution function of $(2T - \sum a_j)/(\sum a_j^2)^{1/2}$ both under the hypothesis and the alternative, an expansion for the power is now immediate.

It remains to discuss the choice of the scores $a_j = a_{j,N}$. For a comparison between best rank tests and best parametric tests we choose a distribution function F with a symmetric smooth density f and consider the locally most powerful (LMP) rank test based on the scores

$$a_{j,N} = E\Psi(U_{j:N}) \quad \text{where } \Psi(t) = -f'F^{-1}((1+t)/2)/fF^{-1}((1+t)/2) \tag{1.10}$$

and $U_{j:N}$ denotes the j-th order statistic of a sample of size N from the uniform distribution on $(0,1)$. Since $F^{-1}((1+t)/2)$ is the inverse function of the distribution function $(2F-1)$ on $(0,\infty)$, $F^{-1}((1+U_{j:N})/2)$ is distributed as the j-th order statistic V_j of the absolute values $|X_1|,|X_2|,\cdots,|X_N|$ of a sample $X_1,X_2,\cdots,X_N$ from F. Hence $a_j = -Ef'(V_j)/f(V_j)$. As f is symmetric, the function f'/f can only be constant on the positive half-line if f is the density $f(x) = 1/2\gamma e^{-\gamma|x|}$ of a Laplace distribution on R^1 for which the sign test is the LMP rank test. We already concluded that this test can not be handled with the tools of this paper, but for every other symmetric four times differentiable f, the important condition (1.9) will hold.

If, instead of the so-called exact scores $a_{j,N} = E\Psi(U_{j:N})$, one uses the approximate scores $a_{j,N} = \Psi(j/(N+1))$, then the power expansions remain unchanged. This is generally not the case for other score generating functions than Ψ.

The most powerful parametric test for the null-hypothesis F against the contiguous shift alternative $F(x-\theta)$ with $\theta = cN^{1/2}$ for fixed $c>0$ will serve as a basis for comparison of the LMP rank test. Its test statistic is simply $\sum_{1\le j\le N}\{\log f(X_j-\theta) - \log f(X_j)\}$ which is a sum of i.i.d. random variables and therefore its distribution function under the hypothesis and the alternative admit Edgeworth expansions under the usual assumptions, and so does the power. Explicit expressions are found for the deficiency of the LMP rank test and some examples are:

Normal distribution (Hodges-Lehmann problem). For normal location alternatives the one-sample normal scores test as well as van der Waerden's one-sample rank test with respect to the most powerful parametric test based on the sample mean equals

$$d_N = 1/2\log\log N + 1/2(u_\alpha^2 - 1) + 1/2\gamma + o(1),$$

where $\Phi(u_\alpha) = 1-\alpha$ and $\gamma = 0.577216$ is Euler's constant. Note that in the paper there is an error in the constant (cf. Albers et al. 1978). In this case the deficiency does tend to infinity, but no one is likely to notice as $1/2\log\log N = 1.568\cdots$ for $N = 10^{10}$ (logarithms to base e).

It is also shown that the deficiency of the permutation test based on the sample mean with respect to Student's one-sample test tends to zero as $O(N^{-1/2})$.

Logistic distribution. For logistic location alternatives the deficiency of Wilcoxon's one-sample test with respect to the most powerful test for testing $F(x) = (1+e^{-x})^{-1}$ against $F(x-bN^{-1/2})$ tends to a finite limit and equals

$$d_N = \{18 + 12u_\alpha^2 + (48)^{1/2}bu_\alpha + b^2\}/60 + o(1).$$

It came as somewhat of a surprise that Wilcoxon's test statistic admits a three-term Edgeworth expansion, as it is a purely lattice random variable. As we pointed out above, the reason that this is possible is that its conditional distribution is that of a sum of independent but not identically distributed random variables. Intuitively the reason is that the point probabilities of the Wilcoxon statistic are of the order $N^{-3/2}$ which is allowed as the error of the expansion is $o(N^{-1})$.

The final section of the paper discusses deficiencies of estimators of location. It is shown that the deficiency of the Hodges-Lehmann type of location estimator associated with the LMP rank test for location alternatives with respect to the maximum likelihood estimator for location, differs by $O(N^{-1/4})$ from the deficiency of the parent tests.

The paper deals with a technically highly complicated subject and is therefore not easy to read. At the time of appearance it had the dubious distinction of being the second longest paper published in the Annals. With 49 pages it was second only to Larry Brown's 50 pages on the admissibility of invariant estimators (Brown 1966). However, for those interested in expansions and higher order asymptotics it contains a veritable treasure of technical achievements that improve our understanding of asymptotic statistics. I hope this review will facilitate the reading. While I'm about it, let me also recommend reading the companion paper (Bickel and van Zwet 1978) where the same program is carried out for two-sample rank tests. With its 68 pages it was regrettably the longest paper in the Annals at the time it was published, but don't let that deter you! Understanding the technical tricks in this area will come in handy in all sorts of applications.

1.1.3 Edgeworth Expansions in Nonparametric Statistics

This paper is a very readable review of the state of the art at the time in the area of Edgeworth expansions. It discusses the extension of Cramér's work to sums of i.i.d. random vectors, as well as expansions for M-estimators. It also gives a preview of the results of the paper we have just discussed on one-sample rank tests and the paper we just mentioned on two-sample rank tests. There is also a new result of Bickel on U-statistics that may be viewed as the precursor of a move towards a general theory of expansions for functions of independent random variables. As we have already discussed Cramér's work as well as rank statistics, let me restrict the discussion of the present paper to the result on U-statistics.

First of all, recall the classical Berry-Esseen inequality for normalized sums $S_N = N^{-1/2} \cdot \sum_{1 \le j \le N} X_j$ of i.i.d. random variables $X_1, \cdots, X_N$, with $EX_1 = 0$ and $EX_1^2 = 1$. If $E|X_1|^3 < \infty$, and Φ denotes the standard normal distribution function, then there exists a constant C such that for all N,

$$\sup_x |P(S_N \le x) - \Phi(x)| \le CE|X_1|^3 N^{-1/2}. \tag{1.11}$$

In the present paper a bound of Berry-Esseen-type is proved for U-statistics. Let $X_1, X_2, \cdots$ be i.i.d. random variables with a common distribution function F and let ψ be a measurable, real-valued function on R^2 where it is bounded, say $|\psi| \le M < \infty$, and symmetric, i.e. $\psi(x,y) = \psi(y,x)$. Define

$$\gamma(x) = E(\psi(X_1, X_2)|X_1 = x) =!_{(0,1)} \psi(x,y)\, dF(y)$$

and suppose that $E\psi(X_1, X_2) = E\gamma(X_1) = 0$. Define a normalized U-statistic T_N by

$$T_N = \sigma_N^{-1} \sum_{1 \le i < j \le N} \psi(X_i, X_j) \quad \text{with } \sigma_N^2 = E\{\sum_{1 \le i < j \le N} \psi(X_i, X_j)\}^2, \tag{1.12}$$

and hence $ET_N = 0$ and $ET_N^2 = 1$. In the paper it is proved that if $E\gamma^2(X_1) > 0$, then there exists a constant C depending on ψ but not on N such that

$$\sup_x |P(T_N \le x) - \Phi(x)| \le CN^{-1/2}. \tag{1.13}$$

When comparing this result with the Berry-Esseen bound for the normalized sum S_N, one gets the feeling that the assumption that ψ is bounded is perhaps a bit too restrictive and that it should be possible to replace it by one or more moment conditions. But it was a good start and improvements were made in quick succession. The boundedness assumption for ψ was dropped and Chan and Wierman (1977) proved the result under the conditions that $E\gamma^2(X_1) > 0$ and $E\{\psi(X_1, X_2)\}^4 < \infty$. Next Callaert and Janssen (1978) showed that $E\gamma^2(X_1) > 0$ and $E|\psi(X_1, X_2)|^3 < \infty$ suffice. Finally Helmers and van Zwet (1982) proved the bound under the assumptions $E\gamma^2(X_1) > 0$, $E|\gamma(X_1)|^3 < \infty$ and $E\psi(X_1, X_2)^2 < \infty$.

Why is this development of interest? The U-statistics discussed so far are a special case of U-statistics of order k which are of the form

$$T = \sum_{\substack{1 \le j(1) < j(2) < \\ \cdots < j(k) \le N}} \psi_k(X_{j(1)}, X_{j(2)}, \cdots, X_{j(k)}), \tag{1.14}$$

where ψ_k is a symmetric function of k variables with $E\psi_k(X_1, X_2, \cdots, X_k) = 0$ and the summation is over all distinct k-tuples chosen from $X_1, X_2, \cdots, X_N$. Clearly the U-statistics discussed above have degree $k = 2$, but extension of the Berry-Esseen inequality to U-statistics of fixed finite degree k is straightforward. In an unpublished technical report (Hoeffding 1961) Wassily Hoeffding showed that any symmetric function $T = t(X_1, \cdots, X_N)$ of N i.i.d. random variables $X_1, \cdots, X_N$ that has $ET = 0$ and finite variance $\sigma^2 = ET^2 - \{ET\}^2 < \infty$ can be written as a sum of U-statistics of orders $k = 1, 2, \cdots, N$ in such a way that all terms involved in this decomposition are uncorrelated and have several additional desirable properties. Hence it seems that it might be possible to obtain results for symmetric functions of N i.i.d. random variables through a study of U-statistics. For the Berry-Esseen theorem this was done in van Zwet (1984) where the result was obtained under fairly mild moment conditions that reduce to the best conditions for U-statistics when specialized to this case. A first step for obtaining Edgeworth expansions for symmetric functions of i.i.d. random variables was taken in Bickel et al. (1986) where the case of U-statistics of degree $k = 2$ was treated. More work is needed here.

References

Albers W, Bickel PJ, van Zwet WR (1976) Asymptotic expansions for the power of distribution free tests in the one-sample problem. Ann Stat 4:108–156

Albers W, Bickel PJ, van Zwet WR (1978) Correction to asymptotic expansions for the power of distributionfree tests in the one-sample problem. Ann Stat 6:1170

Bickel PJ (1974) Edgeworth expansions in nonparametric statistics. Ann Stat 2:1–20

Bickel PJ, van Zwet WR (1978) Asymptotic expansions for the power of distribution free tests in the two-sample problem. Ann Stat 6:937–1004

Bickel PJ, Goetze F, van Zwet WR (1986) The Edgeworth expansion for U-statistics of degree two. Ann Stat 14:1463–1484

Brown LD (1966) On the admissibility of invariant estimators of one or more location parameters. Ann Math Stat 37:1087–1136

Callaert H, Janssen P (1978) The Berry-Esseen theorem for U-statistics. Ann Stat 6:417–421

Chan YK, Wierman J (1977) The Berry-Esseen theorem for U-statistics. Ann Probab 5:136–139

Chebyshev PL (1890) Sur deux théorèmes relatifs aux probabilités. Acta Math 14:305–315

Cramér H (1937) Random variables and probability distributions. Cambridge tracts in mathematics, vol 36. The University press, Cambridge

Edgeworth FY (1905) The law of error. Proc Camb Philos Soc 20:36–65

Feller W (1966) An introduction to probability theory and its applications, vol II. Wiley, New York

Helmers R, van Zwet WR (1982) The Berry-Esseen bound for U-statistics. In: Gupta SS, Berger JO (eds) Statistical decision theory and related topics III, vol 1. Academic, New York, pp 497–512

Hodges JL Jr, Lehmann EL (1970) Deficiency. Ann Math Stat 41:783–801

Hoeffding W (1961) The strong law of large numbers for U-statistics. Institute of statistics, University of North Carolina mimeograph series, vol 302

Stigler SM (1980) Stiglers law of eponymy. Trans New York Acad Sci 39:147–158

van Zwet WR (1984) A Berry-Esseen bound for symmetric statistics. Z. Wahrscheinlichkeitstheorie verw Gebiete 66:425–440

Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics Bulletin 1(6):80–83

The Annals of Statistics
1976, Vol. 4, No. 1, 108–156

ASYMPTOTIC EXPANSIONS FOR THE POWER OF DISTRIBUTION FREE TESTS IN THE ONE-SAMPLE PROBLEM[1]

BY W. ALBERS, P. J. BICKEL[2] AND W. R. VAN ZWET[3]

University of Leiden and University of California, Berkeley

Asymptotic expansions are established for the power of distribution free tests in the one-sample problem. These expansions are then used to obtain deficiencies in the sense of Hodges and Lehmann (1970) for distribution free tests with respect to their parametric competitors and for the estimators of location associated with these tests.

1. Introduction. Let $X_1, \cdots, X_N$ be independent and identically distributed random variables with a common absolutely continuous distribution. For $N = 1, 2, \cdots$, consider the problem of testing the hypothesis that this distribution is symmetric about zero against a sequence of alternatives that is contiguous to the hypothesis as $N \to \infty$. The level α of the sequence of tests is fixed in (0, 1). Standard tests for this problem are linear rank tests and linear permutation tests and expressions for the limiting powers of such tests are of course well-known. In this paper we shall be concerned with obtaining asymptotic expansions to order N^{-1} for the powers π_N of these tests, i.e. expressions of the form $\pi_N = c_0 + c_1 N^{-\frac{1}{2}} + c_{2,N} N^{-1} + o(N^{-1})$. Of course this involves establishing similar expansions for the distribution function of the test statistic under the hypothesis as well as under contiguous alternatives. For simplicity we shall eventually limit our discussion to contiguous location alternatives and in this case terms of order $N^{-\frac{1}{2}}$ do not occur in the expansions.

One reason to consider these problems would be to obtain better numerical approximations for the critical value of the test statistic and the power of the test than can be provided by the usual normal approximation. A number of authors have investigated this possibility, usually dealing only with the hypothesis in order to obtain critical values and more often for the two-sample case than for the one-sample tests we are concerned with here. For an account of this work we refer to a review paper of Bickel (1974), which incidentally also contains a preview of the present study. Here we merely note that, with the exception of a recent paper of Rogers (1971), all previous work is based on formal Edgeworth

Received July 1974; revised June 1975.

[1] Report SW 30/74 Mathematisch Centrum, Amsterdam.

[2] Research supported by the National Science Foundation, Grant GP-38485 A1, and by the Office of Naval Research, Contract N0014-69-A-0200-1038.

[3] Research supported by the National Science Foundation, Grants GP-29123 and GP-31091X, and by the Office of Naval Research, Contract N00014-69-A-0200-1036.

AMS 1970 *subject classifications.* Primary 62G10, 62G20; Secondary 60F05.

Key words and phrases. Distribution free tests, linear rank tests, permutation tests, power, contiguous alternatives, Edgeworth expansions, deficiency.

expansions. One of the purposes of the present paper is to give a rigorous proof of the validity of such expansions. Rogers (1971) has given such a proof for the two-sample Wilcoxon test under the hypothesis. In a companion paper (Bickel and van Zwet (1975)) expansions will be derived for the general two-sample linear rank test under the hypothesis as well as under contiguous location alternatives.

Here we shall not dwell on the numerical aspects of the expansions we obtain. Numerical results are contained in the Ph. D. thesis of Albers (1974). We only mention that the expansions for the power seem to behave as might be expected. In those cases where the normal approximation already produces reasonably good results, the expansions perform even better and often much better. On the other hand, in cases where the normal approximation is known to be disastrous—the Wilcoxon test for Cauchy alternatives for instance—the expansion is as bad or even worse.

We shall concentrate on a different aspect of the expansions for the power. Consider two sequences of tests $\{T_N\}$ and $\{T_N'\}$ for the same hypothesis at the same fixed level α. Let $\pi_N(\theta_N)$ and $\pi_N'(\theta_N)$ denote the powers of these tests against the same sequence of contiguous alternatives parametrized by a parameter θ. If T_N is more powerful than T_N' we search for a number $k_N = N + d_N$ such that $\pi_N(\theta_N) = \pi'_{k_N}(\theta_N)$. Here k_N and d_N are treated as continuous variables, the power π_N' being defined for real N by linear interpolation between consecutive integers. The quantity d_N was named the deficiency of $\{T_N'\}$ with respect to T_N by Hodges and Lehmann (1970), who introduced this concept and initiated its study. Of course, in many cases of interest, d_N is analytically intractable and one can only study its asymptotic behavior as N tends to infinity.

Suppose that for $N \to \infty$, the ratio N/k_N tends to a limit e, the asymptotic relative efficiency of $\{T_N'\}$ with respect to $\{T_N\}$. If $0 < e < 1$, we have $d_N \sim (e^{-1} - 1)N$ and further asymptotic information about d_N is not particularly revealing. On the other hand, if $e = 1$, the asymptotic behavior of d_N, which may now be anything from $o(1)$ to $o(N)$, does provide important additional information. Of special interest is the case where d_N tends to a finite limit, the asymptotic deficiency of $\{T_N'\}$ with respect to $\{T_N\}$ (cf. Hodges and Lehmann (1970)).

Of course, an asymptotic evaluation of d_N is a more delicate matter than showing that $e = 1$. What is needed is an expansion for the power of the type we discussed above. With the aid of such expansions we arrive at the following results. Let F be a distribution function with a density f that is symmetric about zero and let b be a positive real number. Consider the problem of testing the hypothesis F against the sequence of alternatives $F(x - bN^{-\frac{1}{2}})$ at level α. Let d_N denote the deficiency of the locally most powerful rank test with respect to the most powerful test for this problem. Under certain regularity conditions on F we establish an expression for d_N with remainder $o(1)$ and show that this expression remains unchanged if the exact scores in the locally most powerful rank test are replaced by the corresponding approximate scores. The asymptotic behavior of d_N is found to be governed by that of

$$I_N = \int_{1/N}^{1-1/N} \left(\frac{d^2}{dt^2} f\left(F^{-1}\left(\frac{1+t}{2}\right)\right)\right)^2 t(1-t)\, dt \tag{1.1}$$

in the sense that $d_N = O(I_N)$ as $N \to \infty$. By taking F to be the normal distribution we find that the deficiency of both Fraser's normal scores test and van der Waerden's test with respect to the $\bar{X}$-test for contiguous normal alternatives tends to ∞ at the rate of $\frac{1}{2}\log\log N$. For logistic alternatives the deficiency of Wilcoxon's signed rank test with respect to the most powerful parametric test tends to a finite limit. Another typical result is that for contiguous normal alternatives the deficiency of the permutation test based on $\sum X_i$ with respect to Student's test tends to zero for $N \to \infty$.

Combining numerical and Monte Carlo methods, Albers (1974) has evaluated the deficiency of the normal scores test with respect to the $\bar{X}$-test for $N = 5 - (1) - 10$, 20 and 50. The results agree reasonably well with the asymptotic expression for d_N.

To every linear rank test with nonnegative and nondecreasing scores, there corresponds an estimator of location due to Hodges and Lehmann (1963). A similar correspondence exists between the locally most powerful parametric test and the maximum likelihood estimator. We shall exploit this correspondence to obtain asymptotic expansions for the distribution functions of these estimators. We shall show that, when suitably defined, the deficiency of the Hodges–Lehmann estimator associated with the locally most powerful rank test with respect to the maximum likelihood estimator is asymptotically equivalent to the deficiency of the parent tests.

In Section 2 we establish an asymptotic expansion for the distribution function of the general linear rank statistic for the one-sample problem under the hypothesis as well as under alternatives. We specialize to contiguous location alternatives in Section 3 and derive an expansion for the power of the linear rank test. In Section 4 we deal with the important case where the scores are exact or approximate scores generated by a smooth function J. Linear permutation tests are discussed in Section 5. The results on deficiencies of distribution free tests are contained in Section 6. Finally, Section 7 is devoted to estimators.

Although the basic ideas underlying this paper are simple, the proofs are a highly technical matter. The most laborious parts are dealt with in two appendices. We have omitted the proofs of Theorem 5.1 and Lemma 6.1 because we felt that their inclusion would entail much repetition without essentially new ideas. Some relevant results have been left out altogether for much the same reasons. We are referring to a treatment of contiguous alternatives other than location alternatives for linear rank tests, to expansions for the power of locally most powerful parametric tests, most powerful permutation tests and randomized rank score tests. These missing parts may all be found in the Ph. D. thesis of Albers (1974).

2. The basic expansion. Let $X_1, \cdots, X_N$ be independent and identically distributed (i.i.d.) random variables (rv's) with common distribution (df) G and density g, and let $0 < Z_1 < Z_2 < \cdots < Z_N$ denote the order statistics of the absolute values of $X_1, \cdots, X_N$. If $|X_{R_j}| = Z_j$, define

$$V_j = 1 \quad \text{if} \quad X_{R_j} > 0 \qquad (2.1)$$
$$= 0 \quad \text{otherwise.}$$

We introduce a vector of scores $a = (a_1, \cdots, a_N)$ and define the statistic

$$T = \sum_{j=1}^N a_j V_j \, . \qquad (2.2)$$

We shall be concerned with obtaining an asymptotic expansion for the distribution of T as $N \to \infty$.

Our notation strongly suggests that we are considering a fixed underlying df G and perhaps also a fixed infinite sequence of scores as $N \to \infty$. However, this is merely a matter of notational convenience and our main concern will in fact be the case where the df depends on N and the scores form a triangular array $a_{j,N}, j = 1, \cdots, N, N = 1, 2, \cdots$. Since we are suppressing the index N throughout our notation we shall formally present our results in terms of error bounds for a fixed, but arbitrary, value of N. However, as we shall point out following the proof of Theorem 2.2, these results are really asymptotic expansions in disguise.

The rv T is of course the general linear rank statistic for testing the hypothesis that g is symmetric about zero. Under this hypothesis, $V_1, \cdots, V_N$ are i.i.d. with $P(V_j = 1) = \frac{1}{2}$. For general G, $V_1, \cdots, V_N$ are not independent. However, one easily verifies that, conditional on $Z = (Z_1, \cdots, Z_N)$, the rv's $V_1, \cdots, V_N$ are independent with

$$P_j = P(V_j = 1 \mid Z) = \frac{g(Z_j)}{g(Z_j) + g(-Z_j)} \, . \qquad (2.3)$$

As independence allows us to obtain expansions of Edgeworth type, we shall carry out the following program to arrive at an expansion for the distribution of T. First we obtain an Edgeworth expansion for the distribution of $\sum a_j W_j$, where $W_1, \cdots, W_N$ are independent with $p_j = P(W_j = 1) = 1 - P(W_j = 0)$. Having done this we substitute the random vector $P = (P_1, \cdots, P_N)$ defined in (2.3) for $p = (p_1, \cdots, p_N)$ in this expansion. The expected value of the resulting expression will then give us an expansion for the distribution of T.

In carrying out the first part of this program we shall indicate any dependence on $p = (p_1, \cdots, p_N)$ in our notation. Consider the rv

$$\frac{\sum_{j=1}^N a_j(W_j - p_j)}{\tau(p)} \, , \qquad (2.4)$$

where

$$\tau^2(p) = \sum_{j=1}^N p_j(1 - p_j) a_j^2 \qquad (2.5)$$

denotes the variance of $\sum a_j W_j$. Obviously (2.4) has expectation 0 and variance 1; its third and fourth cumulants, multiplied by $N^{\frac{1}{2}}$ and N respectively, are

$$\kappa_3(p) = -N^{\frac{1}{2}} \frac{\sum p_j(1-p_j)(2p_j-1)a_j^3}{\tau^3(p)}, \tag{2.6}$$

$$\kappa_4(p) = N \frac{\sum p_j(1-p_j)(1-6p_j+6p_j^2)a_j^4}{\tau^4(p)}. \tag{2.7}$$

Let R and ρ denote the df and the characteristic function (ch.f.) of (2.4), thus

$$R(x,p) = P\left(\frac{\sum a_j(W_j - p_j)}{\tau(p)} \leqq x\right), \tag{2.8}$$

$$\rho(t,p) = \prod_{j=1}^N \left[p_j \exp\left\{i(1-p_j)\frac{a_j t}{\tau(p)}\right\} + (1-p_j)\exp\left\{-ip_j \frac{a_j t}{\tau(p)}\right\}\right]. \tag{2.9}$$

A formal Edgeworth expansion to order N^{-1} for the df R is given by (Cramér (1946), page 229)

$$\tilde{R}(x,p) = \Phi(x) + \phi(x)\{N^{-\frac{1}{2}}Q_1(x,p) + N^{-1}Q_2(x,p)\}, \tag{2.10}$$

where Φ and ϕ denote the df and the density of the standard normal distribution, and

$$Q_1(x,p) = -\frac{\kappa_3(p)}{6}(x^2-1), \tag{2.11}$$

$$Q_2(x,p) = -\frac{\kappa_4(p)}{24}(x^3-3x) - \frac{\kappa_3^2(p)}{72}(x^5-10x^3+15x).$$

Let $\tilde{r}(x,p)$ be the derivative of $\tilde{R}(x,p)$ with respect to x. In what follows we shall need an expression for the Fourier transform $\tilde{\rho}(t,p) = \int \exp(itx)\tilde{r}(x,p)\,dx$ of $\tilde{r}$ and one easily verifies that

$$\tilde{\rho}(t,p) = e^{-\frac{1}{2}t^2}\left\{1 - \frac{\kappa_3(p)it^3}{6N^{\frac{1}{2}}} + \frac{3\kappa_4(p)t^4 - \kappa_3^2(p)t^6}{72N}\right\}. \tag{2.12}$$

To justify a formal Edgeworth expansion like (2.10), i.e. to show that $|\tilde{R} - R|$ is indeed $o(N^{-1})$, one usually invokes the following result (Feller (1966), page 512).

LEMMA 2.1. *Let R be a* df *with vanishing expectation and* ch.f. ρ. *Suppose that $R - \tilde{R}$ vanishes at $\pm\infty$ and that $\tilde{R}$ has a derivative $\tilde{r}$ such that $|\tilde{r}| \leqq m$. Finally, suppose that $\tilde{r}$ has a continuously differentiable Fourier transform $\tilde{\rho}$ such that $\tilde{\rho}(0) = 1$ and $\tilde{\rho}'(0) = 0$. Then for all x and $T > 0$,*

$$|R(x) - \tilde{R}(x)| \leqq \frac{1}{\pi}\int_{-T}^{T}\left|\frac{\rho(t)-\tilde{\rho}(t)}{t}\right|dt + \frac{24m}{\pi T}. \tag{2.13}$$

To prove that $|R - \tilde{R}| = o(N^{-1})$, it therefore suffices to show that e.g. for $T = bN^{\frac{3}{2}}$, the integral in (2.13) is $o(N^{-1})$. For the case we are considering this may be done in the standard manner (Feller (1966), Chapter 16) with one important modification at the point where it is shown that $|\rho(t,p)/t|$ is sufficiently small when $|t|$ is of the order $\tau(p)$ or larger. Here one usually makes what Feller calls the extravagantly luxurious assumption that the ch.f.'s of all summands are uniformly bounded away from 1 in absolute value outside every neighborhood of 0. Obviously this condition is not satisfied in our case where the summands $a_j W_j$ are lattice rv's. Weaker sufficient conditions of this type are known, but all seem to imply at the very least that the sum itself is nonlattice. In our case this would exclude for instance both the sign test and the Wilcoxon test.

Although the assumptions mentioned above may be unnecessarily strong, it is clear that one has to exclude cases where the sum (2.4) can only assume relatively few different values. As $\tilde{R}$ is continuous, one can not allow R to have jumps of order N^{-1} or larger. Thus the sign test where jumps of order $N^{-\frac{1}{2}}$ occur, will certainly have to be excluded. However, it is exactly the simple lattice character of this statistic that makes it easily amenable to other methods of expansion (see for instance Albers (1974)). For the Wilcoxon statistic on the other hand, all jumps are $O(N^{-\frac{3}{2}})$ and the assumptions we shall make will not rule out this case.

For $0 < \varepsilon < \frac{1}{2}$ and $\zeta > 0$ consider the set of those a_j for which the corresponding p_j satisfies $\varepsilon \leqq p_j \leqq 1 - \varepsilon$, and let $\gamma(\varepsilon, \zeta, p)$ denote the Lebesgue measure λ of the ζ-neighborhood of this set, thus

$$\gamma(\varepsilon, \zeta, p) = \lambda\{x \mid \exists_j |x - a_j| < \zeta, \varepsilon \leqq p_j \leqq 1 - \varepsilon\} . \tag{2.14}$$

LEMMA 2.2. *Suppose that positive numbers c, C, δ and ε exist such that*

$$\frac{1}{N} \sum_{j=1}^N p_j(1 - p_j)a_j^2 \geqq c , \qquad \frac{1}{N} \sum_{j=1}^N a_j^4 \leqq C , \tag{2.15}$$

$$\gamma(\varepsilon, \zeta, p) \geqq \delta N \zeta \qquad \textit{for some} \quad \zeta \geqq N^{-\frac{3}{2}} \log N . \tag{2.16}$$

Then there exist positive numbers b, B and β depending on N, a and p only through c, C, δ and ε, such that

$$\int_{\log(N+1) \leqq |t| \leqq bN^{\frac{3}{2}}} \left| \frac{\rho(t, p) - \tilde{\rho}(t, p)}{t} \right| dt \leqq BN^{-\beta \log N} .$$

PROOF. Since (2.15) implies that $|\kappa_3(p)| \leqq (Cc^{-2})^{\frac{3}{4}}$ and $|\kappa_4(p)| \leqq Cc^{-2}$,

$$\int_{|t| \geqq \log(N+1)} \left| \frac{\tilde{\rho}(t, p)}{t} \right| dt \leqq B_1 N^{-\beta_1 \log N} ,$$

where B_1, $\beta_1 > 0$ depend only on c and C. Also, for all t,

$$\begin{aligned} |\rho(t, p)| &= \prod_{j=1}^N \left\{1 - 2p_j(1 - p_j)\left(1 - \cos \frac{a_j t}{\tau(p)}\right)\right\}^{\frac{1}{2}} \\ &\leqq \exp\left\{-\sum p_j(1 - p_j)\left[\frac{1}{2}\left(\frac{a_j t}{\tau(p)}\right)^2 - \frac{1}{24}\left(\frac{a_j t}{\tau(p)}\right)^4\right]\right\} \\ &\leqq \exp\left\{-\tfrac{1}{2}t^2 + \frac{Ct^4}{96c^2N}\right\} . \end{aligned} \tag{2.17}$$

For $|t| \leqq 4cC^{-\frac{1}{2}}N^{\frac{1}{2}}$ this is $\leqq \exp(-t^2/3)$. Hence, if $b' = 4cC^{-\frac{1}{2}}$, there exist positive constants B_2 and β_2 such that

$$\int_{\log(N+1)\leqq|t|\leqq b'N^{\frac{1}{2}}} \left|\frac{\rho(t,p)}{t}\right| dt \leqq B_2 N^{-\beta_2 \log N}\,.$$

As $\gamma(\varepsilon, \zeta, p)/\zeta$ is nonincreasing in ζ, we may assume that $\zeta \leqq 1$ in (2.16). Because of (2.15), for any $M > \zeta$ the number of $|a_j| \geqq M - \zeta$ can be at most $CN(M-\zeta)^{-4}$; choosing $M = (8C/\delta)^{\frac{1}{4}} + 1$ we have $CN(M-\zeta)^{-4} \leqq \delta N/8 \leqq \gamma(\varepsilon, \zeta, p)/8\zeta$. It follows that

$$\lambda\{x \mid \exists_j |a_j| \geqq M - \zeta,\ |x - a_j| < \zeta\} \leqq 2\zeta\,\frac{\gamma(\varepsilon, \zeta, p)}{8\zeta} = \frac{\gamma(\varepsilon, \zeta, p)}{4}\,.$$

Together with (2.16) this implies that for every real t

$$\lambda\left\{z \mid \exists_j |a_j| \leqq M - \zeta, \left|z - \frac{a_j t}{\tau(p)}\right| < \frac{\zeta|t|}{\tau(p)},\ \varepsilon \leqq p_j \leqq 1 - \varepsilon\right\} \geqq \frac{3|t|\gamma(\varepsilon, \zeta, p)}{4\tau(p)}\,.$$

Take $b = \delta[(32M/\pi c^{\frac{1}{2}}) + (16/b')]^{-1}$. Then, for every $|t| \in [b'N^{\frac{1}{2}}, bN^{\frac{3}{2}}]$

$$\begin{aligned}
&\lambda\left\{z \,\Big|\, |z| \leqq \frac{M|t|}{\tau(p)},\ |z - k\pi| \leqq \frac{2\zeta b N^{\frac{3}{2}}}{\tau(p)} \text{ for some integer } k\right\} \\
&\quad \leqq \left(\frac{2M|t|}{\pi\tau(p)} + 1\right)\frac{4\zeta b N^{\frac{3}{2}}}{\tau(p)} \leqq \left(\frac{2M|t|}{\pi(cN)^{\frac{1}{2}}} + \frac{|t|}{b'N^{\frac{1}{2}}}\right)\frac{4bN^{\frac{3}{2}}}{\tau(p)}\,\frac{\gamma(\varepsilon, \zeta, p)}{\delta N} = \frac{|t|\gamma(\varepsilon, \zeta, p)}{4\tau(p)}\,,
\end{aligned}$$

and hence

$$\begin{aligned}
\lambda\Bigg\{z \,\Big|\, |z| \leqq \frac{M|t|}{\tau(p)},\ &\exists_j |a_j| \leqq M - \zeta, \left|z - \frac{a_j t}{\tau(p)}\right| < \frac{\zeta|t|}{\tau(p)},\ \varepsilon \leqq p_j \leqq 1 - \varepsilon; \\
&|z - k\pi| > \frac{2\zeta b N^{\frac{3}{2}}}{\tau(p)} \text{ for every integer } k\Bigg\} \geqq \frac{|t|\gamma(\varepsilon, \zeta, p)}{2\tau(p)}\,.
\end{aligned}$$

As $\zeta|t| \leqq \zeta b N^{\frac{3}{2}}$, this implies that the number of indices j for which $|(a_j t/\tau(p)) - k\pi| > \zeta b N^{\frac{3}{2}}/\tau(p)$ for every integer k and $\varepsilon \leqq p_j \leqq 1 - \varepsilon$, is at least equal to

$$\frac{\tau(p)}{2\zeta|t|}\cdot\frac{|t|\gamma(\varepsilon, \zeta, p)}{2\tau(p)} \geqq \frac{\delta N}{4}\,.$$

For such an index j we have for all $|t| \in [b'N^{\frac{1}{2}}, bN^{\frac{3}{2}}]$,

$$\begin{aligned}
\left\{1 - 2p_j(1-p_j)\left(1 - \cos\frac{a_j t}{\tau(p)}\right)\right\}^{\frac{1}{2}} &\leqq \left\{1 - 2\varepsilon(1-\varepsilon)\frac{\zeta^2 b^2 N^3}{(\pi\tau(p))^2}\right\}^{\frac{1}{2}} \\
&\leqq \exp\left\{-\frac{\varepsilon(1-\varepsilon)\zeta^2 b^2 N^3}{(\pi\tau(p))^2}\right\}
\end{aligned}$$

and hence, as $4\tau^2(p) \leqq C^{\frac{1}{2}}N$ and $\zeta \geqq N^{-\frac{3}{2}}\log N$,

$$|\rho(t,p)| \leqq \exp\left\{-\frac{\delta\varepsilon(1-\varepsilon)b^2N^4\zeta^2}{4\pi^2\tau^2(p)}\right\} \leqq \exp\left\{-\frac{\delta\varepsilon(1-\varepsilon)b^2}{\pi^2 C^{\frac{1}{2}}}(\log N)^2\right\}.$$

This implies that for some B_3, $\beta_3 > 0$ depending on c, C, δ and ε,

$$\int_{b'N^{\frac{1}{2}} \leqq |t| \leqq bN^{\frac{3}{2}}} \left| \frac{\rho(t, p)}{t} \right| dt \leqq B_3 N^{-\beta_3 \log N} ,$$

which completes the proof. □

We now justify expansion (2.10).

THEOREM 2.1. *Suppose that positive numbers c, C, δ and ε exist such that* (2.15) *and* (2.16) *are satisfied. Then there exists $A > 0$ depending on N, a and p only through c, C, δ and ε such that*

$$\sup_x |R(x, p) - \tilde{R}(x, p)| \leqq AN^{-\frac{5}{4}} . \tag{2.18}$$

PROOF. For $0 \leqq y \leqq 1$ and $-\pi/2 \leqq z \leqq \pi/2$, $\mathrm{Re}\,[y \exp\{i(1 - y)z\} + (1 - y) \exp\{-iyz\}] \geqq \frac{1}{2}$, and hence we have the following Taylor expansion (mod. $2\pi i$)

$$\begin{aligned} \log (ye^{i(1-y)z} &+ (1 - y)e^{-iyz}) \\ &= -\tfrac{1}{2}y(1 - y)z^2 + \tfrac{1}{6}y(1 - y)(2y - 1)iz^3 \\ &\quad + \tfrac{1}{24}y(1 - y)(1 - 6y + 6y^2)z^4 + M_1(y, z) , \end{aligned} \tag{2.19}$$

where $|M_1(y, z)| \leqq C_1|z|^5$ for some fixed $C_1 > 0$. If $|a_j t/\tau(p)| \leqq \pi/2$ for all j, we can apply this expansion to the logarithm of every factor in (2.9) which yields

$$\rho(t, p) = \exp \left\{ -\tfrac{1}{2}t^2 - \frac{\kappa_3(p)it^3}{6N^{\frac{1}{2}}} + \frac{\kappa_4(p)t^4}{24N} + M_2(t, p) \right\} , \tag{2.20}$$

where $|M_2(t, p)| \leqq C_1|t/\tau(p)|^5 \sum |a_j|^5$.

Condition (2.15) implies that $\max |a_j| \leqq (CN)^{\frac{1}{4}}$ and hence that $|a_j t/\tau(p)| \leqq (Cc^{-2})^{\frac{1}{4}}N^{-\frac{1}{4}}|t|$ for all j. We have already seen that $|\kappa_3(p)| \leqq (Cc^{-2})^{\frac{3}{4}}$ and $|\kappa_4(p)| \leqq Cc^{-2}$; because $\max |a_j| \leqq (CN)^{\frac{1}{4}}$ we also have $\tau^{-5}(p) \sum |a_j|^5 \leqq (Cc^{-2})^{\frac{5}{4}}N^{-\frac{5}{4}}$. It follows from these remarks that there exists $c_1 > 0$, depending only on c and C, such that for $|t| \leqq c_1 N^{\frac{1}{4}}$ expansion (2.20) is valid and also

$$\left| -\frac{\kappa_3(p)it^3}{6N^{\frac{1}{2}}} \right| + \left| \frac{\kappa_4(p)t^4}{24N} \right| + |M_2(t, p)| \leqq \tfrac{1}{4}t^2 .$$

Hence, for $|t| \leqq c_1 N^{\frac{1}{4}}$, Taylor expansion of (2.20) yields

$$\rho(t, p) = \tilde{\rho}(t, p) + M_3(t, p) , \tag{2.21}$$

where $\tilde{\rho}$ is given by (2.12), $|M_3(t, p)| \leqq (N^{-\frac{3}{2}} + N^{-\frac{5}{2}} \sum |a_j|^5)|t|^5 Q(|t|) \exp(-t^2/4)$, and Q is a polynomial with coefficients depending on c and C. This implies the existence of $A_1 > 0$ depending on c and C and such that

$$\int_{|t| \leqq c_1 N^{\frac{1}{4}}} \left| \frac{\rho(t, p) - \tilde{\rho}(t, p)}{t} \right| dt \leqq A_1 N^{-\frac{5}{4}} . \tag{2.22}$$

As c_1 depends only on c and C we may assume without loss of generality that N is so large that $\log (N + 1) \leqq c_1 N^{\frac{1}{4}}$. The theorem is now proved by combining

(2.22) and Lemma 2.2, noting that $\tilde{r}(x, t) = (\partial/\partial x)\tilde{R}(x, t)$ is bounded by a number depending only on c and C and applying Lemma 2.1. □

It will be clear that by requiring that $\sum |a_j|^5 \leqq CN$ in Theorem 2.1 one obtains $|R - \tilde{R}| \leqq AN^{-\frac{3}{2}}$ which is the "natural" order of the remainder.

Before we replace p by the random vector $P = (P_1, \cdots, P_N)$ defined in (2.3) and compute the unconditional distribution of T by taking the expected value, we first have to change the standardization of $\sum a_j W_j$ into one that does not involve p. As before, let $W_1, \cdots, W_N$ be independent with $P(W_j = 1) = 1 - P(W_j = 0) = p_j$, let $\bar{p} = (\bar{p}_1, \cdots, \bar{p}_N)$ be a vector with $0 \leqq \bar{p}_j \leqq 1$ for all j, and consider the df $R^*(x, p, \bar{p})$ of the rv $\tau^{-1}(\bar{p}) \sum a_j(W_j - \bar{p}_j)$, thus

$$R^*(x, p, \bar{p}) = P\left(\frac{\sum a_j(W_j - \bar{p}_j)}{\tau(\bar{p})} \leqq x\right). \tag{2.23}$$

Here $\tau^2(\bar{p}) = \sum \bar{p}_j(1 - \bar{p}_j)a_j^2$ in accordance with (2.5); similarly $\kappa_3(\bar{p})$, $\kappa_4(\bar{p})$, $Q_1(x, \bar{p})$, $Q_2(x, \bar{p})$ and $\tilde{R}(x, \bar{p})$ are defined by replacing p by $\bar{p}$ in (2.6), (2.7), (2.11) and (2.10).

For reasons that will become clear in the sequel we shall also at this stage expand $\tau(\bar{p})/\tau(p)$ in powers of $(\tau^2(p) - \tau^2(\bar{p}))/\tau^2(\bar{p})$; at the same time the numerators of $\kappa_3(p)$ and $\kappa_4(p)$ will be expanded about the point $p = \bar{p}$. Later on, when p_j is replaced by P_j, we shall e.g. take $\bar{p}_j = EP_j$ thus ensuring that $P_j - \bar{p}_j$ is roughly speaking a rv of order $N^{-\frac{1}{2}}$. At the moment, however, we do not make any assumptions about $p - \bar{p}$ and as a result Lemma 2.3 provides only a formal expansion in the sense that we do not claim that the remainder term is at all small.

The expansion for $R^*(x, p, \bar{p})$ that we shall establish is

$$\begin{aligned} \tilde{R}^*(x, p, \bar{p}) = \tilde{R}(x - u, \bar{p}) - \phi(x - u) \Bigg\{ & \frac{1}{2} \frac{\tau^2(p) - \tau^2(\bar{p})}{\tau^2(\bar{p})} (x - u) \\ & + \frac{1}{6} \frac{\sum (p_j - \bar{p}_j)(1 - 6\bar{p}_j + 6\bar{p}_j^2)a_j^3}{\tau^3(\bar{p})} [(x - u)^2 - 1] \\ & + \frac{1}{8} \left(\frac{\tau^2(p) - \tau^2(\bar{p})}{\tau^2(\bar{p})}\right)^2 [(x - u)^3 - 3(x - u)] \\ & + \frac{\kappa_3(\bar{p})}{12N^{\frac{1}{2}}} \frac{\tau^2(p) - \tau^2(\bar{p})}{\tau^2(\bar{p})} [(x - u)^4 - 6(x - u)^2 + 3] \Bigg\}, \end{aligned} \tag{2.24}$$

where $\tilde{R}$ is given by (2.10) and

$$u = \frac{\sum (p_j - \bar{p}_j)a_j}{\tau(\bar{p})}. \tag{2.25}$$

LEMMA 2.3. *Let $\bar{p} = (\bar{p}_1, \cdots, \bar{p}_N)$ be a vector of real numbers in* $[0, 1]$ *and suppose that positive numbers c, C, δ and ε exist such that* (2.15) *and* (2.16) *are satisfied and that*

$$\frac{1}{N} \sum_{j=1}^N \bar{p}_j(1 - \bar{p}_j)a_j^2 \geqq c. \tag{2.26}$$

Then there exists $A > 0$ depending on N, a, p and $\bar{p}$ only through c, C, δ and ε and such that

$$\text{(2.27)}\qquad \sup_x |R^*(x, p, \bar{p}) - \tilde{R}^*(x, p, \bar{p})| \leqq A\{N^{-\frac{5}{4}} + N^{-\frac{3}{2}} \textstyle\sum (p_j - \bar{p}_j)^2|a_j|^3 + N^{-3}|\tau^2(p) - \tau^2(\bar{p})|^3\} .$$

PROOF. Changing the standardization in Theorem 2.1 we find

$$\text{(2.28)}\qquad \sup_x \left|R^*(x, p, \bar{p}) - \tilde{R}\left((x-u)\frac{\tau(\bar{p})}{\tau(p)}, p\right)\right| \leqq AN^{-\frac{5}{4}} .$$

The assumptions of the lemma ensure that $\tau^2(\bar{p})/\tau^2(p) \geqq cC^{-\frac{1}{2}}$, $\tau^2(p)/\tau^2(\bar{p}) \geqq cC^{-\frac{1}{2}}$, $|\kappa_3(p)| \leqq (c^{-2}C)^{\frac{3}{4}}$, $|\kappa_3(\bar{p})| \leqq (c^{-2}C)^{\frac{3}{4}}$, $|\kappa_4(p)| \leqq c^{-2}C$ and $|\kappa_4(\bar{p})| \leqq c^{-2}C$. It follows that the derivatives of $\tilde{R}((x-u)y, p)$ with respect to y are bounded for $y^2 \geqq cC^{-\frac{1}{2}}$ and all $x - u$, and hence

$$\text{(2.29)}\qquad \begin{aligned} &\tilde{R}\left((x-u)\frac{\tau(\bar{p})}{\tau(p)}, p\right) \\ &= \tilde{R}(x-u, p) + \tilde{R}'(x-u, p)\left(\frac{\tau(\bar{p})}{\tau(p)} - 1\right)(x-u) \\ &\quad + \tfrac{1}{2}\tilde{R}''(x-u, p)\left(\frac{\tau(\bar{p})}{\tau(p)} - 1\right)^2(x-u)^2 + O\left(\left(\frac{\tau(\bar{p})}{\tau(p)} - 1\right)^3\right), \end{aligned}$$

where $\tilde{R}'(x, p)$ and $\tilde{R}''(x, p)$ denote first and second derivatives of $\tilde{R}(x, p)$ with respect to x. Since $(\tau^2(p) - \tau^2(\bar{p}))/\tau^2(\bar{p}) \geqq -1 + cC^{-\frac{1}{2}}$,

$$\text{(2.30)}\qquad \frac{\tau(\bar{p})}{\tau(p)} = 1 - \frac{1}{2}\,\frac{\tau^2(p) - \tau^2(\bar{p})}{\tau^2(\bar{p})} + \frac{3}{8}\left(\frac{\tau^2(p) - \tau^2(\bar{p})}{\tau^2(\bar{p})}\right)^2 - \cdots ,$$

where the remainder is of the order of the first term omitted. As $\kappa_3(\bar{p})$ and $\kappa_4(\bar{p})$ are bounded, we obtain the following one and two term expansions with remainder for $\kappa_3(p)$ and $\kappa_4(p)$.

$$\text{(2.31)}\qquad \begin{aligned} \kappa_3(p) &= \left[\kappa_3(\bar{p}) - N^{\frac{1}{2}}\frac{\sum \{p_j(1-p_j)(2p_j-1) - \bar{p}_j(1-\bar{p}_j)(2\bar{p}_j-1)\}a_j^3}{\tau^3(\bar{p})}\right]\left(\frac{\tau(\bar{p})}{\tau(p)}\right)^3 \\ &= \kappa_3(\bar{p}) + O(N^{-1}|\tau^2(p) - \tau^2(\bar{p})| + N^{-1}\textstyle\sum |p_j - \bar{p}_j|\,|a_j|^3) \\ &= \kappa_3(\bar{p})\left[1 - \frac{3}{2}\,\frac{\tau^2(p) - \tau^2(\bar{p})}{\tau^2(\bar{p})}\right] + N^{\frac{1}{2}}\frac{\sum (p_j - \bar{p}_j)(1 - 6\bar{p}_j + 6\bar{p}_j^2)a_j^3}{\tau^3(\bar{p})} \\ &\quad + O(N^{-2}(\tau^2(p) - \tau^2(\bar{p}))^2 + N^{-1}\textstyle\sum (p_j - \bar{p}_j)^2|a_j|^3 \\ &\quad + N^{-2}|\tau^2(p) - \tau^2(\bar{p})|\textstyle\sum |p_j - \bar{p}_j|\,|a_j|^3) , \end{aligned}$$

$$\text{(2.32)}\qquad \kappa_4(p) = \kappa_4(\bar{p}) + O(N^{-1}|\tau^2(p) - \tau^2(\bar{p})| + N^{-1}\textstyle\sum |p_j - \bar{p}_j|a_j^4) .$$

In (2.29) we may now replace $\tilde{R}$, $\tilde{R}'$ and $\tilde{R}''$ by explicit expressions and substitute (2.32) and appropriate versions of (2.31) and (2.30). The algebra is straightforward and will be omitted. Combining the result with (2.28) we find that (2.27) holds if a term

$$O(N^{-2}\textstyle\sum |p_j - \bar{p}_j|(|a_j|^3 + a_j^4) + N^{-\frac{5}{2}}|\tau^2(p) - \tau^2(\bar{p})|\sum |p_j - \bar{p}_j|\,|a_j|^3 \\ + N^{-2}|\tau^2(p) - \tau^2(\bar{p})| + N^{-\frac{5}{2}}(\tau^2(p) - \tau^2(\bar{p}))^2)$$

is added to the right-hand side. Here, as well as above, the order symbol is uniform for fixed c and C. The lemma is now proved by noting that

$$
\begin{aligned}
&N^{-2} \sum |p_j - \tilde{p}_j|\,|a_j|^3 \leqq N^{-\frac{5}{2}} \sum |a_j|^3 + N^{-\frac{3}{2}} \sum (p_j - \tilde{p}_j)^2 |a_j|^3 ,\\
&N^{-2} \sum |p_j - \tilde{p}_j| a_j^4 \leqq N^{-\frac{5}{2}} \sum |a_j|^5 + N^{-\frac{3}{2}} \sum (p_j - \tilde{p}_j)^2 |a_j|^3 ,\\
&N^{-\frac{5}{2}} |\tau^2(p) - \tau^2(\tilde{p})| \sum |p_j - \tilde{p}_j|\,|a_j|^3 \leqq N^{-\frac{3}{2}} \sum (p_j - \tilde{p}_j)^2 |a_j|^3 \\
&\qquad\qquad\qquad\qquad + N^{-\frac{7}{2}} (\tau^2(p) - \tau^2(\tilde{p}))^2 \sum |a_j|^3 ,\\
&N^{-2} |\tau^2(p) - \tau^2(\tilde{p})| + N^{-\frac{5}{2}} (\tau^2(p) - \tau^2(\tilde{p}))^2 \leqq N^{-\frac{3}{2}} + N^{-3} |\tau^2(p) - \tau^2(\tilde{p})|^3 ,
\end{aligned}
$$

and that $\sum |a_j|^3 \leqq C^{\frac{3}{4}} N$ and $\sum |a_j|^5 \leqq (CN)^{\frac{5}{4}}$. □

We shall now replace p by $P = (P_1, \cdots, P_N)$ in $\tilde{R}^*(x, p, \tilde{p})$ and take expectations. Define the vector $\pi = (\pi_1, \cdots, \pi_N)$ by

$$\pi_j = EP_j , \qquad j = 1, \cdots, N; \tag{2.33}$$

it will play the role of $\tilde{p}$. Furthermore, for $\zeta > 0$ we let $\gamma(\zeta)$ denote the Lebesgue measure λ of the ζ-neighborhood of the set $\{a_1, \cdots, a_N\}$, thus

$$\gamma(\zeta) = \lambda\{x \mid \exists_j |x - a_j| < \zeta\} . \tag{2.34}$$

THEOREM 2.2. *Let $X_1, \cdots, X_N$ be* i.i.d. *with common* df G *and density g, and let T, P and π be defined by* (2.2), (2.3) *and* (2.33). *Suppose that positive numbers c, C, δ, δ' and ε exist with $\delta' <$* min $(\delta/2, c^2 C^{-1})$ *and such that*

$$\frac{1}{N} \sum_{j=1}^N a_j^2 \geqq c , \qquad \frac{1}{N} \sum_{j=1}^N a_j^4 \leqq C , \tag{2.35}$$

$$\gamma(\zeta) \geqq \delta N \zeta \qquad \textit{for some} \quad \zeta \geqq N^{-\frac{3}{4}} \log N , \tag{2.36}$$

$$P\left(\varepsilon \leqq \frac{g(X_1)}{g(X_1) + g(-X_1)} \leqq 1 - \varepsilon\right) \geqq 1 - \delta' . \tag{2.37}$$

Then there exists $A > 0$ depending on N, a and G only through c, C, δ, δ' and ε, and such that

$$
\begin{aligned}
\sup_x \left| P\left(\frac{T - \sum a_j \pi_j}{\tau(\pi)} \leqq x\right) - E\tilde{R}^*(x, P, \pi)\right| \\
\leqq A\{N^{-\frac{5}{4}} + N^{-\frac{3}{4}}[\sum \{E(P_j - \pi_j)^2\}^{\frac{5}{2}}]^{\frac{2}{5}} + N^{-\frac{3}{2}}[\sum \{E|P_j - \pi_j|^3\}^{\frac{2}{3}}]^{\frac{3}{2}}\} .
\end{aligned}
\tag{2.38}
$$

PROOF. We start by showing that a, P and π satisfy the conditions for a, p and $\tilde{p}$ in Lemma 2.3 with large probability.

The number of P_j that lie in $[\varepsilon, 1 - \varepsilon]$ is equal to the number of $g(X_j)/(g(X_j) + g(-X_j))$ in that interval. Applying an exponential bound for binomial probabilities (Okamoto (1958)) we find that for $\delta'' \in (\delta', \min(\delta/2, c^2C^{-1}))$, (2.37) implies

$$P(\varepsilon \leqq P_j \leqq 1 - \varepsilon \;\text{ for at least }\; (1 - \delta'')N \;\text{ indices }\; j) \geqq 1 - e^{-2N(\delta'' - \delta')^2} .$$

Suppose that $\varepsilon \leqq P_j \leqq 1 - \varepsilon$ for at least $(1 - \delta'')N$ values of j. It then follows from (2.36) that a and P satisfy condition (2.16) if δ is replaced by $\delta - 2\delta'' > 0$.

For $\eta \in (0, 1)$, suppose that $a_j^2 \leqq \eta c$ for exactly k indices j and let $\sum'$ indicate summation over the remaining $N - k$ indices. Because of (2.35)

$$c \leqq \frac{1}{N} \sum a_j^2 \leqq \frac{k}{N} \eta c + \frac{1}{N} \sum' a_j^2 \leqq \eta c + \frac{N-k}{N} \left(\frac{1}{N-k} \sum' a_j^4 \right)^{\frac{1}{2}}$$
$$\leqq \eta c + \left(\frac{N-k}{N} C \right)^{\frac{1}{2}},$$

and hence the number of $a_j^2 > \eta c$ is at least $(1-\eta)^2 c^2 C^{-1} N$. By choosing η sufficiently small we can ensure that $(1-\eta)^2 c^2 C^{-1} > \delta''$. This implies that $N^{-1}\tau^2(P) \geqq \tilde{c}$, where $\tilde{c} = ((1-\eta)^2 c^2 C^{-1} - \delta'')\varepsilon(1-\varepsilon)\eta c > 0$. This in turn ensures that $N^{-1}\tau^2(\pi) \geqq N^{-1} E\tau^2(P) \geqq c^*$, where $c^* = \tilde{c}(1 - \exp\{-2(\delta'' - \delta')^2\}) > 0$.

Thus we have shown that if c, C, δ and ε are replaced by positive numbers c^*, C, $\delta - 2\delta''$ and ε depending only on c, C, δ, δ' and ε, then a and π satisfy (2.26) and the second part of (2.15), whereas a and P satisfy (2.16) and the first part of (2.15) except on a set E with $P(E) \leqq \exp\{-2N(\delta'' - \delta')^2\} = O(N^{-\frac{5}{4}})$. Hence a, P and π satisfy the assumptions of Lemma 2.3 on the complement of E. In dealing with the set E it will suffice to note that $\tilde{R}^*(x, P, \pi)$ is bounded since (2.26) and the second part of (2.15) ensure the boundedness of $\kappa_3(\pi)$, $\kappa_4(\pi)$, $(\tau^2(P) - \tau^2(\pi))/\tau^2(\pi)$ and $\sum |a_j|^3/\tau^3(\pi)$. Of course $R^*(x, P, \pi)$, being a probability, is also bounded.

As

$$P\left(\frac{T - \sum a_j \pi_j}{\tau(\pi)} \leqq x \right) = ER^*(x, P, \pi),$$

the left-hand side of (2.38) is bounded above by

$$E \sup_x |R^*(x, P, \pi) - \tilde{R}^*(x, P, \pi)|. \tag{2.39}$$

Applying Lemma 2.3 on the complement of E and using the boundedness of $|R^*(x, P, \pi) - \tilde{R}^*(x, P, \pi)|$ together with $P(E) = O(N^{-\frac{5}{4}})$ we find that (2.39) is

$$O(N^{-\frac{5}{4}} + N^{-\frac{5}{2}} \sum E(P_j - \pi_j)^2 |a_j|^3 + N^{-3} E|\tau^2(P) - \tau^2(\pi)|^3),$$

where the order symbol is uniform for fixed c, C, δ, δ' and ε. Now

$$N^{-\frac{5}{2}} \sum E(P_j - \pi_j)^2 |a_j|^3 \leqq N^{-\frac{5}{2}} [\sum \{E(P_j - \pi_j)^2\}^{\frac{5}{2}}]^{\frac{2}{5}} (\sum |a_j|^5)^{\frac{3}{5}},$$
$$N^{-3} E|\tau^2(P) - \tau^2(\pi)|^3 \leqq N^{-3} E[\sum |P_j - \pi_j| a_j^2]^3 \leqq N^{-3} [\sum \{E|P_j - \pi_j|^3\}^{\frac{1}{3}} a_j^2]^3$$
$$\leqq N^{-3} [\sum \{E|P_j - \pi_j|^3\}^{\frac{2}{3}}]^{\frac{3}{2}} (\sum a_j^4)^{\frac{3}{2}},$$

and since $\sum |a_j|^5 \leqq (CN)^{\frac{5}{4}}$ and $\sum a_j^4 \leqq CN$, this completes the proof. □

We note that the boundedness of $\tilde{R}^*(x, P, \pi)$ on E plays an important role in the above proof. Because $\tau(P)$ may be arbitrarily small on E, this explains why we had to remove $\tau(p)$ from the denominator of the expansion in Lemma 2.3 by means of (2.30).

Although Theorem 2.2 is formally stated as a result for a fixed, but arbitrary value of N, it is of course meaningless for fixed N because we do not investigate the way in which A depends on c, C, δ, δ' and ε. In fact the theorem is a purely asymptotic result. Let us for a moment indicate dependence on N by a superscript. Thus, for $N = 1, 2, \cdots$, consider the distribution of the statistic $T^{(N)}$ based on a vector of scores $a^{(N)} = (a_1^{(N)}, \cdots, a_N^{(N)})$ when the underlying df is $G^{(N)}$. Fix positive values of c, C, δ, δ' and ε with $\delta' < \min(\delta/2, c^2C^{-1})$. The theorem asserts that if for every N, $a^{(N)}$ and $G^{(N)}$ satisfy (2.35)—(2.37) for these fixed c, C, δ, δ' and ε, then the error of the approximation $E\tilde{R}^*(x, P^{(N)}, \pi^{(N)})$ is

$$O(N^{-\frac{5}{4}} + N^{-\frac{3}{4}}[\sum \{E(P_j^{(N)} - \pi_j^{(N)})^2\}^{\frac{5}{2}}]^{\frac{2}{5}} + N^{-\frac{3}{2}}[\sum \{E|P_j^{(N)} - \pi_j^{(N)}|^3\}^{\frac{2}{3}}]^{\frac{3}{2}})$$

as $N \to \infty$. Moreover, the order of the remainder is uniform for all such sequences $a^{(N)}$, $G^{(N)}$, $N = 1, 2, \cdots$.

Assumption (2.36) may need some clarification. It is clear from the proof of Lemma 2.2 that the role of conditions (2.16) and (2.36) in Theorems 2.1 and 2.2 is to ensure that the a_j do not cluster too much around too few points. Assumption (2.36) is certainly satisfied if for some $k \geqq \delta N/2$, indices $j_1, j_2, \cdots, j_k$ exist such that $a_{j_{i+1}} - a_{j_i} \geqq 2N^{-\frac{3}{2}} \log N$ for $i = 1, \cdots, k-1$. Under condition (2.35) this will typically be the case. Consider for instance the important case $a_j = EJ(U_{j:N})$, where $U_{1:N} < U_{2:N} < \cdots < U_{N:N}$ are order statistics from the uniform distribution on (0, 1) and J is a continuously differentiable, nonconstant function on (0, 1) with $\int J^4 < \infty$. Here both (2.35) and (2.36) are satisfied for all N with fixed c, C and δ. The same is true if $a_j = J(j/(N+1))$ provided that J is monotone near 0 and 1.

For a large class of underlying df's G, the right-hand side of (2.38) is uniformly $o(N^{-1})$. Still Theorem 2.2 does not yet provide an explicit expansion to order N^{-1} for the distribution of T since we are still left with the task of computing the expected value of $\tilde{R}^*(x, P, \pi)$. This is of course a trivial matter under the hypothesis that g is symmetric about zero and, more generally, in the case where, for some $\eta > 0$, $g(x)/g(-x) = \eta$ for all $x > 0$. In this case $P_j = \eta(1+\eta)^{-1}$ with probability 1 for all j and an expansion for the distribution of T is already contained in Theorem 2.1. For fixed alternatives in general, however, the computation of $E\tilde{R}^*(x, P, \pi)$ presents a formidable problem that we shall not attempt to solve here. It would seem that what is needed, is an expansion for the distribution of a linear combination of functions of order statistics.

In the remaining part of this paper we shall restrict attention to sequences of alternatives that are contiguous to the hypothesis. Heuristically the situation is now as follows. Since $g(x)/(g(x) + g(-x)) = \frac{1}{2} + O(N^{-\frac{1}{2}})$, $P_j - \frac{1}{2}$ and $\pi_j - \frac{1}{2}$ will be $O(N^{-\frac{1}{2}})$, whereas $P_j - \pi_j$ will be $O(N^{-1})$ instead of $O(N^{-\frac{1}{2}})$ as before. In the first place this allows us to simplify $E\tilde{R}^*(x, P, \pi)$ considerably as a number of terms may now be relegated to the remainder and functions of π_j may be expanded about the point $\pi_j = \frac{1}{2}$. Much more important, however, is the fact that $U^* = \tau^{-1}(\pi) \sum (P_j - \pi_j)a_j$ will now be $O(N^{-\frac{1}{2}})$ and that we may therefore expand $\tilde{R}^*(x, P, \pi)$ in powers of U^*. This means that we shall be dealing with low moments of linear combinations of functions of order statistics rather than with their distributions. We need hardly point out that a heuristic argument like this can be entirely misleading and that the actual order of the remainder in our expansion will of course have to be investigated. The unduly complicated form of the remainder terms in the preceeding theorem is, of course, preparatory to such further expansion.

Define

$$\tilde{K}(x) = \Phi(x) + \phi(x)\left\{\frac{\sum a_j^2 E(2P_j - 1)^2 - 4\sigma^2(\sum a_j P_j)}{2\sum a_j^2}\, x + \frac{\sum a_j^3(2\pi_j - 1)}{3(\sum a_j^2)^{\frac{3}{2}}}(x^2 - 1) + \frac{\sum a_j^4}{12(\sum a_j^2)^2}(x^3 - 3x)\right\}, \tag{2.40}$$

where $\sigma^2(Z)$ denotes the variance of a rv Z. Carrying out the type of computation outlined above we arrive at the following simplified version of Theorem 2.2.

THEOREM 2.3. *Theorem* 2.2 *continues to hold if* (2.38) *is replaced by*

$$\begin{aligned}\sup_x\left|P\left(\frac{2T - \sum a_j}{(\sum a_j^2)^{\frac{1}{2}}} \leqq x\right) - \tilde{K}\left(x - \frac{\sum a_j(2\pi_j - 1)}{(\sum a_j^2)^{\frac{1}{2}}}\right)\right| \\ \leqq A\{N^{-\frac{5}{4}} + \sum\{E(2P_j - 1)^4\}^{\frac{5}{4}} + N^{-\frac{3}{4}}[\sum\{E|P_j - \pi_j|^3\}^{\frac{4}{9}}]^{\frac{9}{4}}\}\,.\end{aligned} \tag{2.41}$$

PROOF. The proof of this theorem becomes somewhat shorter if we use a modification of Theorem 2.2 as a starting point rather than Theorem 2.2 itself. We recall that Theorem 2.2 was proved by an application of Lemma 2.3 for $\bar{p} = \pi$. However, the proof clearly goes through for any other choice of $\bar{p}$ that satisfies (2.26). Because of (2.35), we may therefore replace π in (2.38) by a vector $\bar{p}$ with $\bar{p}_j = \frac{1}{2}$ for all j. Noting that for this choice of $\bar{p}$, $\kappa_3(\bar{p}) = 0$, $\kappa_4(\bar{p}) = -2N\sum a_j^4/(\sum a_j^2)^2$, $\tau^2(P) - \tau^2(\bar{p}) = -\frac{1}{4}\sum(2P_j - 1)^2 a_j^2$, and adding the last two terms in $\tilde{R}^*(x, P, \bar{p})$ to the remainder, we obtain

$$\begin{aligned}P&\left(\frac{2T - \sum a_j}{(\sum a_j^2)^{\frac{1}{2}}} \leqq x\right) \\ &= E\Phi(x - \tilde{U}) + E\phi(x - \tilde{U})\left\{\frac{\sum a_j^4}{12(\sum a_j^2)^2}[(x - \tilde{U})^3 - 3(x - \tilde{U})]\right. \\ &\quad + \frac{\sum a_j^2(2P_j - 1)^2}{2\sum a_j^2}(x - \tilde{U}) \\ &\quad \left. + \frac{\sum a_j^3(2P_j - 1)}{3(\sum a_j^2)^{\frac{3}{2}}}[(x - \tilde{U})^2 - 1]\right\} \\ &\quad + O(N^{-\frac{5}{4}} + N^{-\frac{3}{4}}[\sum\{E(2P_j - 1)^2\}^{\frac{5}{2}}]^{\frac{2}{5}} + N^{-\frac{3}{2}}[\sum\{E|2P_j - 1|^3\}^{\frac{2}{3}}]^{\frac{3}{2}} \\ &\quad + N^{-2}E[\sum a_j^2(2P_j - 1)^2]^2 + N^{-\frac{3}{2}}\sum a_j^2 E(2P_j - 1)^2)\,,\end{aligned} \tag{2.42}$$

where $\tilde{U} = \sum a_j(2P_j - 1)/(\sum a_j^2)^{\frac{1}{2}}$. All order symbols in this proof are uniform for fixed c, C, δ, δ' and ε. The remainder in (2.42) may be simplified by noting that

$$\begin{aligned}N^{-\frac{3}{4}}&[\sum\{E(2P_j - 1)^2\}^{\frac{5}{2}}]^{\frac{2}{5}} + N^{-\frac{3}{2}}[\sum\{E|2P_j - 1|^3\}^{\frac{2}{3}}]^{\frac{3}{2}} \\ &\leqq N^{-\frac{5}{4}} + \sum\{E(2P_j - 1)^2\}^{\frac{5}{2}} + N^{-1}\sum E|2P_j - 1|^3 \\ &\leqq N^{-\frac{5}{4}} + N^{-\frac{3}{2}} + 2\sum\{E(2P_j - 1)^4\}^{\frac{5}{4}}\,,\end{aligned}$$

$$\begin{aligned}
N^{-\frac{3}{2}}E[\sum a_j^2(2P_j-1)^2]^2 + N^{-\frac{3}{2}}\sum a_j^2E(2P_j-1)^2
&\leqq 2N^{-\frac{3}{2}}E[\sum a_j^2(2P_j-1)^2]^2 + N^{-\frac{3}{2}} \\
&\leqq 2N^{-\frac{3}{2}}\sum a_j^4\sum E(2P_j-1)^4 + N^{-\frac{3}{2}} \\
&\leqq 2C\sum \{E(2P_j-1)^4\}^{\frac{5}{4}} + (2C+1)N^{-\frac{3}{2}} .
\end{aligned}$$

Define $U = \sum a_j(P_j - \pi_j)/(\sum a_j^2)^{\frac{1}{2}}$, so $x - \tilde{U} = x - \sum a_j(2\pi_j - 1)/(\sum a_j^2)^{\frac{1}{2}} - 2U$. By expanding in powers of U under the expectation sign in (2.42) we find

$$\begin{aligned}
&P\left(\frac{2T-\sum a_j}{(\sum a_j^2)^{\frac{1}{2}}} \leqq x\right) \\
(2.43)\qquad &= \tilde{K}\left(x - \frac{\sum a_j(2\pi_j-1)}{(\sum a_j^2)^{\frac{1}{2}}}\right) + O(N^{-\frac{5}{4}} + \sum\{E(2P_j-1)^4\}^{\frac{5}{4}} + E|U|^3 \\
&\quad + E|U|\{N^{-1} + N^{-1}\sum a_j^2(2P_j-1)^2 + N^{-\frac{3}{2}}\sum |a_j|^3|2P_j-1|\}) .
\end{aligned}$$

Now

$$\begin{aligned}
N^{-\frac{3}{2}}\sum |a_j|^3|2P_j-1| &\leqq N^{-2}\sum a_j^4 + N^{-1}\sum a_j^2(2P_j-1)^2 , \\
N^{-1}E|U| &\leqq N^{-\frac{3}{2}} + E|U|^3 , \\
N^{-1}E|U|\sum a_j^2(2P_j-1)^2 &\leqq N^{-\frac{1}{2}}EU^2 + N^{-\frac{3}{2}}E[\sum a_j^2(2P_j-1)^2]^2 \\
&\leqq N^{-\frac{3}{2}} + E|U|^3 + C\sum\{E(2P_j-1)^4\}^{\frac{5}{4}} + CN^{-\frac{3}{2}} ,
\end{aligned}$$

where the last inequality is based on a bound obtained earlier in this proof. It follows that the remainder in (2.43) is of the order of the sum of its first three terms. The proof is completed by noting that

$$\begin{aligned}
E|U|^3 &\leqq (cN)^{-\frac{3}{2}}E[\sum |a_j|\,|P_j-\pi_j|]^3 \leqq (cN)^{-\frac{3}{2}}[\sum |a_j|\{E|P_j-\pi_j|^3\}^{\frac{1}{3}}]^3 \\
&\leqq (cN)^{-\frac{3}{2}}(\sum a_j^4)^{\frac{3}{4}}[\sum \{E|P_j-\pi_j|^3\}^{\frac{4}{9}}]^{\frac{9}{4}} . \ \square
\end{aligned}$$

Theorem 2.3 provides the basic expansion for the distribution of T under contiguous alternatives. In Section 3 we shall be concerned with a further simplification of this expansion and a precise evaluation of the order of the remainder term.

3. **Contiguous location alternatives.** The analysis in this section will be carried out for contiguous location alternatives rather than for contiguous alternatives in general. The general case can be treated in much the same way as the location case but the conditions as well as the results become more involved. The interested reader is referred to Albers (1974).

Let F be a df with a density f that is positive on R^1, symmetric about zero and four times differentiable with derivatives $f^{(i)}$, $i = 1, \cdots, 4$. Define functions

$$(3.1)\qquad \psi_i = \frac{f^{(i)}}{f} , \qquad i = 1, \cdots, 4 ,$$

and suppose that positive numbers ε and C exist such that for

$$\begin{aligned}
(3.2)\qquad & m_1 = 6 , \quad m_2 = 3 , \quad m_3 = \tfrac{4}{3} , \quad m_4 = 1 , \\
& \sup\{\textstyle\int_{-\infty}^{\infty} |\psi_i(x+y)|^{m_i} f(x)\,dx : |y| \leqq \varepsilon\} \leqq C , \qquad i = 1, \cdots, 4 .
\end{aligned}$$

Let $X_1, \cdots, X_N$ be i.i.d. with common df $G(x) = F(x - \theta)$ where

$$0 \leqq \theta \leqq CN^{-\frac{1}{2}} \tag{3.3}$$

for some positive C. Note that (3.2) and (3.3) together imply contiguity. Let $0 < Z_1 < Z_2 < \cdots < Z_N$ denote the order statistics of $|X_1|, \cdots, |X_N|$ and let T be defined by (2.2). Probabilities, expected values and variances under G will be denoted by P_θ, E_θ and σ_θ^2; under F they will be indicated by P_0, E_0 and σ_0^2. Define

$$\begin{aligned} K_\theta(x) = \Phi(x) + \phi(x) \Bigg\{ & \frac{\sum a_j^4}{12(\sum a_j^2)^2}(x^3 - 3x) - \theta \frac{\sum a_j^3 E_0 \psi_1(Z_j)}{3(\sum a_j^2)^{\frac{3}{2}}}(x^2 - 1) \\ & + \frac{\theta^2}{2 \sum a_j^2}[\sum a_j^2 E_0 \psi_1^2(Z_j) - \sigma_0^2(\sum a_j \psi_1(Z_j))]x \\ & + \frac{\theta^3}{6(\sum a_j^2)^{\frac{1}{2}}} \sum a_j E_0[3\psi_1^3(Z_j) - 6\psi_1(Z_j)\psi_2(Z_j) + \psi_3(Z_j)] \Bigg\} , \end{aligned} \tag{3.4}$$

and

$$\eta = -\theta \frac{\sum a_j E_0 \psi_1(Z_j)}{(\sum a_j^2)^{\frac{1}{2}}} . \tag{3.5}$$

We shall show that $K_\theta(x - \eta)$ is an expansion to order N^{-1} for the df of $(2T - \sum a_j)/(\sum a_j^2)^{\frac{1}{2}}$. The expansion will be established in Theorem 3.1 and an evaluation of the order of the remainder will be given in Theorem 3.2.

Let $\pi(\theta)$ denote the power of the one-sided level α test based on T for the hypothesis of symmetry against the alternative $G(x) = F(x - \theta)$. Suppose that for some $\varepsilon > 0$,

$$\varepsilon \leqq \alpha \leqq 1 - \varepsilon . \tag{3.6}$$

We prove that an expansion for $\pi(\theta)$ is given by

$$\tilde{\pi}(\theta) = 1 - K_\theta(u_\alpha - \eta) + \phi(u_\alpha - \eta) \frac{\sum a_j^4}{12(\sum a_j^2)^2}(u_\alpha^3 - 3u_\alpha) , \tag{3.7}$$

where $u_\alpha = \Phi^{-1}(1 - \alpha)$ denotes the upper α-point of the standard normal distribution.

THEOREM 3.1. *Suppose that positive numbers, c, C, δ and ε exist such that* (2.35), (2.36), (3.2) *and* (3.3) *are satisfied. Then there exists $A > 0$ depending on N, a, F and θ only through c, C, δ and ε and such that*

$$\begin{aligned} \sup_x \left| P_\theta\left(\frac{2T - \sum a_j}{(\sum a_j^2)^{\frac{1}{2}}} \leqq x \right) - K_\theta(x - \eta) \right| \\ \leqq A\{N^{-\frac{5}{4}} + N^{-\frac{3}{4}}\theta^3[\sum \{E_0|\psi_1(Z_j) - E_0\psi_1(Z_j)|^3\}^{\frac{4}{9}}]^{\frac{9}{4}}\} , \end{aligned} \tag{3.8}$$

$$|\eta| \leqq A , \tag{3.9}$$

$$\begin{aligned} \theta \frac{|\sum a_j^3 E_0 \psi_1(Z_j)|}{(\sum a_j^2)^{\frac{3}{2}}} \leqq AN^{-1} , \qquad \theta^2 \frac{\sum a_j^2 E_0 \psi_1^2(Z_j)}{\sum a_j^2} \leqq AN^{-1} , \\ \frac{\theta^3}{(\sum a_j^2)^{\frac{1}{2}}} |\sum a_j E_0[3\psi_1^3(Z_j) - 6\psi_1(Z_j)\psi_2(Z_j) + \psi_3(Z_j)]| \leqq AN^{-1} . \end{aligned} \tag{3.10}$$

If, in addition, (3.6) *is satisfied there exists $A' > 0$ depending on N, a, F, θ and α only through c, C, δ and ε and such that*

$$|\pi(\theta) - \tilde{\pi}(\theta)| \leqq A'\{N^{-\frac{5}{4}} + N^{-\frac{3}{4}}\theta^3[\sum \{E_0|\psi_1(Z_j) - E_0\psi_1(Z_j)|^3\}^{\frac{4}{9}}]^{\frac{9}{4}}\} . \tag{3.11}$$

PROOF. We begin by checking assumption (2.37). One easily verifies that

$$\left|\frac{\partial}{\partial\theta}\frac{f(x-\theta) - f(x+\theta)}{f(x-\theta) + f(x+\theta)}\right| \leqq \tfrac{1}{2}|\psi_1(x-\theta)| + \tfrac{1}{2}|\psi_1(x+\theta)| .$$

Hence the symmetry of f and an application of Markov's inequality and Fubini's theorem yield

$$\begin{aligned} P_\theta\left(\varepsilon \leqq \frac{g(X_1)}{g(X_1) + g(-X_1)} \leqq 1 - \varepsilon\right) &\\ = P_\theta\left(\left|\frac{f(X_1-\theta) - f(X_1+\theta)}{f(X_1-\theta) + f(X_1+\theta)}\right| \leqq 1 - 2\varepsilon\right) &\\ \geqq P_\theta(\textstyle\int_0^\theta \{|\psi_1(X_1 - t)| + |\psi_1(X_1+t)|\}\, dt \leqq 2(1-2\varepsilon)) &\\ \geqq 1 - \frac{1}{2(1-2\varepsilon)} E_\theta \textstyle\int_0^\theta \{|\psi_1(X_1 - t)| + |\psi_1(X_1+t)|\}\, dt &\\ \geqq 1 - \frac{\theta}{1 - 2\varepsilon} \sup_{|t|\leqq\theta} E_\theta|\psi_1(X_1 + t)| . & \end{aligned}$$

Take $\varepsilon < \frac{1}{2}$ and choose $\delta' = \frac{1}{2}\min(\delta/2, c^2C^{-1})$. Because of (3.3) there exists $N_0 > 0$ depending only on c, C, δ and ε such that for $N \geqq N_0$, $2\theta \leqq \varepsilon$ and $\theta \leqq (1-2\varepsilon)C^{-\frac{1}{6}}\delta'$. Then (3.2) implies that (2.37) is satisfied for $N \geqq N_0$. This is of course sufficient to ensure that the conclusion of Theorem 2.3 holds.

The passage from (2.41) to (3.8) is achieved by Taylor expansion with respect to θ. Since this part of the proof is highly technical and laborious it will not be given in the body of the text. Instead we refer the interested reader to Appendix 1 where the results we shall need are stated in Corollary A1.1. Using parts (A1.27), (A1.31) and (A1.32) of Corollary A1.1 together with the inequality $\sum \{E_\theta(2P_j - 1)^4\}^{\frac{5}{4}} \leqq \sum E_\theta|2P_j - 1|^5$ we see that the left-hand side of (3.8) is bounded by the right-hand side of (3.8) plus a term

$$O(\theta^{\frac{19}{8}}\{E_0|\sum a_j(\psi_1(Z_j) - E_0\psi_1(Z_j))|^3\}^{\frac{1}{3}} + N^{-\frac{1}{2}}\theta^5\sigma_0^2(\sum a_j\psi_1(Z_j))) . \tag{3.12}$$

Here, and later in this proof all order symbols are uniform for fixed c, C, δ and ε. Now

$$\begin{aligned} \theta^{\frac{19}{8}}\{E_0|\sum a_j(\psi_1(Z_j) - E_0\psi_1(Z_j))|^3\}^{\frac{1}{3}} + N^{-\frac{1}{2}}\theta^5\sigma_0^2(\sum a_j\psi_1(Z_j)) &\\ \leqq \theta^{\frac{27}{16}} + \theta^6 E_0|\sum a_j(\psi_1(Z_j) - E_0\psi_1(Z_j))|^3 &\\ + N^{-\frac{1}{2}}\theta^3 + N^{-\frac{1}{2}}\theta^6\sigma_0^3(\sum a_j\psi_1(Z_j)) &\\ = O(N^{-\frac{5}{4}} + N^{-\frac{3}{2}}\theta^3 E_0|\sum a_j(\psi_1(Z_j) - E_0\psi_1(Z_j))|^3) , &\\ E_0|\sum a_j(\psi_1(Z_j) - E_0\psi_1(Z_j))|^3 \leqq [\sum |a_j|\{E_0|\psi_1(Z_j) - E_0\psi_1(Z_j)|^3\}^{\frac{1}{3}}]^3 &\\ \leqq (CN)^{\frac{3}{4}}[\sum \{E_0|\psi_1(Z_j) - E_0\psi_1(Z_j)|^3\}^{\frac{4}{9}}]^{\frac{9}{4}} , & \end{aligned}$$

which proves (3.8). In view of (2.35) and (3.3) it is clear that (3.9) and (3.10) are merely restating parts (A1.28)—(A1.30) of Corollary A1.1.

The one-sided level α test based on T rejects the hypothesis if $(2T - \sum a_j)(\sum a_j^2)^{-\frac{1}{2}} \geqq \xi_\alpha$ with possible randomization if equality occurs. Taking $\theta = 0$ in (3.8) we find that

$$1 - \Phi(\xi_\alpha) - \phi(\xi_\alpha) \frac{\sum a_j^4}{12(\sum a_j^2)^2} (\xi_\alpha^3 - 3\xi_\alpha) = \alpha + O(N^{-\frac{5}{4}}) ,$$

and hence because of (2.35) and (3.6),

$$\xi_\alpha = u_\alpha - \frac{\sum a_j^4}{12(\sum a_j^2)^2} (u_\alpha^3 - 3u_\alpha) + O(N^{-\frac{5}{4}}) . \tag{3.13}$$

The power of this test against the alternative $F(x - \theta)$ is

$$\begin{aligned} \pi(\theta) = {} & 1 - K_\theta(\xi_\alpha - \eta) \\ & + O(N^{-\frac{5}{4}} + N^{-\frac{3}{4}}\theta^3[\textstyle\sum \{E_0|\psi_1(Z_j) - E_0\psi_1(Z_j)|^3\}^{\frac{4}{9}}]^{\frac{9}{4}}) . \end{aligned} \tag{3.14}$$

In (3.14) we expand $K_\theta(\xi_\alpha - \eta)$ around $u_\alpha - \eta$. Noting that $|\xi_\alpha - u_\alpha| = O(N^{-1})$ and using (2.35) and (3.10) we arrive at the conclusion that the left-hand side of (3.11) is bounded by the right-hand side of (3.11) plus a term

$$O(N^{-2}\theta^2\sigma_0^2(\textstyle\sum a_j\psi_1(Z_j))) = O(N^{-3} + N^{-\frac{3}{2}}\theta^3E_0|\sum a_j(\psi_1(Z_j) - E_0\psi_1(Z_j))|^3) .$$

As we have already shown earlier in this proof that such a term does not change the order of the remainder in (3.11), the proof of Theorem 3.1 is completed. □

For $i = 1, 2, 3$, define functions Ψ_i on $(0, 1)$ by

$$\Psi_i(t) = \psi_i\left(F^{-1}\left(\frac{1+t}{2}\right)\right) = \frac{f^{(i)}\left(F^{-1}\left(\frac{1+t}{2}\right)\right)}{f\left(F^{-1}\left(\frac{1+t}{2}\right)\right)} . \tag{3.15}$$

THEOREM 3.2. *Suppose that positive numbers C and δ exist such that* (3.3) *is satisfied and that $|\Psi_1'(t)| \leqq C(t(1-t))^{-\frac{4}{3}+\delta}$ for all $0 < t < 1$. Then there exists $A'' > 0$ depending on N, F and θ only through C and δ and such that*

$$N^{-\frac{3}{4}}\theta^3[\textstyle\sum \{E_0|\psi_1(Z_j) - E_0\psi_1(Z_j)|^3\}^{\frac{4}{9}}]^{\frac{9}{4}} \leqq A''N^{-\frac{5}{4}} .$$

For the highly technical proof of this result the reader is referred to Appendix 2. Theorem 3.2 follows at once from Corollary A2.1 in this appendix by taking $h = \Psi_1$.

4. Exact and approximate scores. The expansions given in Section 3 can be simplified further if we make certain smoothness assumptions about the scores a_j. Consider a continuous function J on (0, 1) and let $U_{1:N} < U_{2:N} < \cdots < U_{N:N}$ denote order statistics of a sample of size N from the uniform distribution on (0, 1). For $N = 1, 2, \cdots$ we define the exact scores generated by J by

$$a_j = a_{j,N} = EJ(U_{j:N}) , \qquad j = 1, \cdots, N , \tag{4.1}$$

and the approximate scores generated by J by

$$a_j = a_{j,N} = J\left(\frac{j}{N+1}\right), \qquad j = 1, \cdots, N. \tag{4.2}$$

For almost all well-known linear rank tests the scores are of one of these two types. The locally most powerful rank test against location alternatives of type F is based on exact scores generated by the function $-\Psi_1$, where Ψ_1 is defined in (3.15).

So far, we have systematically kept the order of the remainder in our expansions down to $O(N^{-\frac{5}{4}})$. From this point on, however, we shall be content with a remainder that is $o(N^{-1})$, because otherwise we would have to impose rather restrictive conditions. In the previous sections we have also consistently stressed the fact that the remainder depends on a and F only through certain constants occurring in our conditions, thus in effect indicating classes of scores and distributions for which the expansion holds uniformly. As the number of these constants is becoming rather large, we prefer to formulate our results from here on for a fixed score function J and a fixed df F. The reader can easily construct uniformity classes for himself by using the results of Section 3 and tracing the development of Appendix 2.

DEFINITION 4.1. $\mathscr{J}$ is the class of functions J on (0, 1) that are twice continuously differentiable and nonconstant on (0, 1), and satisfy

$$\textstyle\int_0^1 J^4(t)\, dt < \infty . \tag{4.3}$$

$$\limsup_{t\to 0,1} t(1-t)\left|\frac{J''(t)}{J'(t)}\right| < \tfrac{3}{2}. \tag{4.4}$$

$\mathscr{F}$ is the class of df's F on R^1 with positive densities f that are symmetric about zero, four times differentiable and such that, for $\psi_i = f^{(i)}/f$, $\Psi_i(t) = \psi_i(F^{-1}((1+t)/2))$, $m_1 = 6$, $m_2 = 3$, $m_3 = \frac{4}{3}$, $m_4 = 1$,

$$\limsup_{y\to 0} \textstyle\int_{-\infty}^{\infty} |\psi_i(x+y)|^{m_i} f(x)\, dx < \infty, \qquad i = 1, \cdots, 4, \tag{4.5}$$

$$\limsup_{t\to 0,1} t(1-t)\left|\frac{\Psi_1''(t)}{\Psi_1'(t)}\right| < \tfrac{3}{2}. \tag{4.6}$$

For $J \in \mathscr{J}$ and $F \in \mathscr{F}$, let

$$\begin{aligned} \tilde{K}_\theta(x) = \Phi(x) + \phi(x) \Big\{ & N^{-1} \frac{\int_0^1 J^4(t)\, dt}{12(\int_0^1 J^2(t)\, dt)^2} (x^3 - 3x) \\ & - N^{-\frac{1}{2}}\theta \frac{\int_0^1 J^3(t)\Psi_1(t)\, dt}{3(\int_0^1 J^2(t)\, dt)^{\frac{3}{2}}} (x^2 - 1) + \frac{\theta^2}{2\int_0^1 J^2(t)\, dt} \\ & \times \left[\textstyle\int_0^1 J^2(t)\Psi_1^2(t)\, dt - \int_0^1\int_0^1 J(s)\Psi_1'(s)J(t)\Psi_1'(t)(s\wedge t - st)\, ds\, dt\right] x \\ & + \frac{N^{\frac{1}{2}}\theta^3}{6(\int_0^1 J^2(t)\, dt)^{\frac{1}{2}}} \textstyle\int_0^1 J(t)[3\Psi_1^3(t) - 6\Psi_1(t)\Psi_2(t) + \Psi_3(t)]\, dt \Big\}, \end{aligned} \tag{4.7}$$

$$K_{\theta,1}(x) = \tilde{K}_\theta(x) + \phi(x)\frac{N^{-\frac{1}{2}}\theta}{2(\int_0^1 J^2(t)\,dt)^{\frac{1}{2}}}\left\{\frac{\int_0^1 J(t)\Psi_1(t)\,dt}{\int_0^1 J^2(t)\,dt}\sum_{j=1}^N \sigma^2(J(U_{j:N})) - 2\sum_{j=1}^N \operatorname{Cov}(J(U_{j:N}), \Psi_1(U_{j:N}))\right\}, \tag{4.8}$$

$$K_{\theta,2}(x) = \tilde{K}_\theta(x) + \phi(x)\frac{N^{-\frac{1}{2}}\theta}{2(\int_0^1 J^2(t)\,dt)^{\frac{1}{2}}}\left\{\frac{\int_0^1 J(t)\Psi_1(t)\,dt}{\int_0^1 J^2(t)\,dt}\int_{1/N}^{1-1/N}(J'(t))^2 t(1-t)\,dt - 2\int_{1/N}^{1-1/N} J'(t)\Psi_1'(t)t(1-t)\,dt\right\}, \tag{4.9}$$

$$\tilde{\eta} = -N^{\frac{1}{2}}\theta\frac{\int_0^1 J(t)\Psi_1(t)\,dt}{(\int_0^1 J^2(t)\,dt)^{\frac{1}{2}}}, \tag{4.10}$$

$$\pi_i(\theta) = 1 - K_{\theta,i}(u_\alpha - \tilde{\eta}) + \phi(u_\alpha - \tilde{\eta})N^{-1}\frac{\int_0^1 J^4(t)\,dt}{12(\int_0^1 J^2(t)\,dt)^2}(u_\alpha^3 - 3u_\alpha), \tag{4.11}$$

for $i = 1, 2$. Then, in the notation of Section 3, we have for contiguous location alternatives and exact scores

THEOREM 4.1. *Let $F \in \mathscr{F}$, $J \in \mathscr{J}$, $a_j = EJ(U_{j:N})$ for $j = 1, \cdots, N$, and let $0 \leqq \theta \leqq CN^{-\frac{1}{2}}$, $\varepsilon \leqq \alpha \leqq 1 - \varepsilon$ for positive C and ε. Then, for every fixed J, F, C and ε, there exist positive numbers A, $\delta_1, \delta_2, \cdots$ such that $\lim_{N\to\infty}\delta_N = 0$ and for every N*

$$\sup_x\left|P_\theta\left(\frac{2T - \sum a_j}{(\sum a_j^2)^{\frac{1}{2}}} \leqq x\right) - K_{\theta,1}(x - \tilde{\eta})\right| \leqq \delta_N N^{-1}, \tag{4.12}$$

$$\sup_x\left|P_\theta\left(\frac{2T - \sum a_j}{(\sum a_j^2)^{\frac{1}{2}}} \leqq x\right) - K_{\theta,2}(x - \tilde{\eta})\right| \leqq \delta_N N^{-1} + AN^{-\frac{3}{2}}\int_{1/N}^{1-1/N}|J'(t)|(|J'(t)| + |\Psi_1'(t)|)(t(1-t))^{\frac{1}{2}}\,dt, \tag{4.13}$$

$$|\pi(\theta) - \pi_1(\theta)| \leqq \delta_N N^{-1} \tag{4.14}$$

$$|\pi(\theta) - \pi_2(\theta)| \leqq \delta_N N^{-1} + AN^{-\frac{3}{2}}\int_{1/N}^{1-1/N}|J'(t)|(|J'(t)| + |\Psi_1'(t)|)(t(1-t))^{\frac{1}{2}}\,dt. \tag{4.15}$$

PROOF. For fixed $J \in \mathscr{J}$, positive constants c, C and δ exist for which (2.35) and (2.36) hold for all N (cf. one of the remarks following the proof of Theorem 2.2). Similarly, for fixed $F \in \mathscr{F}$, (3.2) is satisfied and it follows that the conclusions of Theorem 3.1 hold with A and A' depending only on F, J, C and ε. Also (4.5) ensures that Ψ_1^6 is summable and together with (4.6) and the second part of Corollary A2.1, this implies that the conclusion of Theorem 3.2 holds with A'' depending only on F and C.

To complete the proof we now apply the results collected in Corollary A2.2 to the expansions $K_\theta(x - \eta)$ and $\tilde{\pi}(\theta)$ in Theorem 3.1 and then expand these functions of η around the point $\eta = \tilde{\eta}$, while noting that $\eta - \tilde{\eta} = o(N^{-\frac{1}{2}})$ by (A2.22) and (A2.23). □

In general, the expansions given in Theorem 4.1 will not hold if the exact scores are replaced by approximate scores $a_j = J(j/(N+1))$, because $\eta - \tilde{\eta}$ will then give rise to a different term of order N^{-1}. If $J = -\Psi_1$, however, it is clear from Corollary A2.2 and the proof of Theorem 4.1 that expansions (4.13) and (4.15) are valid for approximate as well as exact scores. Also for $J = -\Psi_1$, these expansions may be simplified because $F \in \mathscr{F}$ implies that by partial integration

$$\int_0^1 \int_0^1 \Psi_1(s)\Psi_1'(s)\Psi_1(t)\Psi_1'(t)(s \wedge t - st)\, ds\, dt = \tfrac{1}{4} \int_0^1 \Psi_1^4(t)\, dt - \tfrac{1}{4}\left(\int_0^1 \Psi_1^2(t)\, dt\right)^2 ,$$
$$\int_0^1 \Psi_1(t)[6\Psi_1(t)\Psi_2(t) - \Psi_3(t)]\, dt = \tfrac{10}{3} \int_0^1 \Psi_1^4(t)\, dt + \int_0^1 \Psi_2^2(t)\, dt .$$

It follows that in this case $\tilde{\eta}$, $K_{\theta,2}(x - \tilde{\eta})$ and $\pi_2(\theta)$ reduce to

$$\eta^* = N^{\frac{1}{2}}\theta\left(\int_0^1 \Psi_1^2(t)\, dt\right)^{\frac{1}{2}} , \tag{4.16}$$

$$\begin{aligned} L_\theta(x) = \Phi(x - \eta^*) &+ \frac{\phi(x - \eta^*)}{72N} \\ &\times \left\{ \frac{\int_0^1 \Psi_1^4(t)\, dt}{(\int_0^1 \Psi_1^2(t)\, dt)^2} [6(x^3 - 3x) + 6\eta^*(x^2 - 1) - 3\eta^{*2}x - 5\eta^{*3}] \right. \\ &+ \frac{12 \int_0^1 \Psi_2^2(t)\, dt}{(\int_0^1 \Psi_1^2(t)\, dt)^2} \eta^{*3} + 9\eta^{*2}(x - \eta^*) \\ &\left. + \frac{36 \int_{1/N}^{1-1/N} (\Psi_1'(t))^2 t(1-t)\, dt}{\int_0^1 \Psi_1^2(t)\, dt} \eta^* \right\} , \end{aligned} \tag{4.17}$$

$$\begin{aligned} \pi^*(\theta) = 1 - \Phi(u_\alpha - \eta^*) &+ \frac{\eta^*\phi(u_\alpha - \eta^*)}{72N} \\ &\times \left\{ \frac{\int_0^1 \Psi_1^4(t)\, dt}{(\int_0^1 \Psi_1^2(t)\, dt)^2} [-6(u_\alpha^2 - 1) + 3\eta^* u_\alpha + 5\eta^{*2}] \right. \\ &- \frac{12 \int_0^1 \Psi_2^2(t)\, dt}{(\int_0^1 \Psi_1^2(t)\, dt)^2} \eta^{*2} - 9\eta^*(u_\alpha - \eta^*) \\ &\left. - \frac{36 \int_{1/N}^{1-1/N} (\Psi_1'(t))^2 t(1-t)\, dt}{\int_0^1 \Psi_1^2(t)\, dt} \right\} . \end{aligned} \tag{4.18}$$

Finally we note that for $F \in \mathscr{F}$, $-\Psi_1$ can not be constant on (0, 1) because the density $f(x) = \frac{1}{2}\lambda e^{-\lambda|x|}$ of the double exponential distribution is not differentiable at zero. It follows that $-\Psi_1 \in \mathscr{J}$ for every $F \in \mathscr{F}$. We have proved

THEOREM 4.2. *Let $F \in \mathscr{F}$ and let either $a_j = -E\Psi_1(U_{j:N})$ for $j = 1, \cdots, N$ or $a_j = -\Psi_1(j/(N+1))$ for $j = 1, \cdots, N$. Suppose that $0 \leqq \theta \leqq CN^{-\frac{1}{2}}$ and $\varepsilon \leqq \alpha \leqq 1 - \varepsilon$ for positive C and ε. Then, for every fixed F, C and ε, there exist positive numbers $A, \delta_1, \delta_2, \cdots$ such that $\lim_{N\to\infty} \delta_N = 0$ and for every N*

$$\sup_x \left| P_\theta\left(\frac{2T - \sum a_j}{(\sum a_j^2)^{\frac{1}{2}}} \leqq x\right) - L_\theta(x) \right| \leqq \delta_N N^{-1} + AN^{-\frac{5}{4}} \int_{1/N}^{1-1/N} (\Psi_1'(t))^2 (t(1-t))^{\frac{1}{2}}\, dt , \tag{4.19}$$

$$|\pi(\theta) - \pi^*(\theta)| \leqq \delta_N N^{-1} + AN^{-\frac{5}{4}} \int_{1/N}^{1-1/N} (\Psi_1'(t))^2 (t(1-t))^{\frac{1}{2}}\, dt . \tag{4.20}$$

At this point it may be useful to make some remarks concerning the assumptions in Theorems 4.1 and 4.2. Conditions (4.4) and (4.6) ensure that J' and Ψ_1' do not oscillate too wildly near 0 and 1. They also limit the growth of these functions near 0 and 1, but in this respect conditions (4.3) and (4.5) for $i = 1$ are typically much stronger. Together with (4.4) and (4.6) they imply that $J'(t) = o((t(1-t))^{-\frac{5}{4}})$ and $\Psi_1'(t) = o((t(1-t))^{-\frac{3}{2}})$ near 0 and 1 (cf. the proof of Corollary A2.1).

For expansions (4.13), (4.15), (4.19) and (4.20) to be meaningful rather than just formally correct, even stronger growth conditions have to be imposed. Consider, for example, expansion (4.20) and suppose, as is typically the case, that Ψ_1' remains bounded near 0. If $\Psi_1'(t) = o((1-t)^{-1})$ near 1, then the right-hand side in (4.20) is $o(N^{-1})$ and the expansion makes sense. However, if $\Psi_1'(t)$ is of exact order $(1-t)^{-1}$, the expansion reduces to

$$\pi(\theta) = 1 - \Phi(u_\alpha - \eta^*) - \frac{\eta^*\phi(u_\alpha - \eta^*)}{2N} \frac{\int_0^{1-1/N} (\Psi_1'(t))^2 t(1-t)\,dt}{\int_0^1 \Psi_1^2(t)\,dt} + O(N^{-1}).$$

Finally, if $\Psi_1'(t) \sim (1-t)^{-1-\delta}$ for $t \to 1$ and some $0 < \delta < \frac{1}{6}$, then all we have left in (4.20) is $\pi(\theta) = 1 - \Phi(u_\alpha - \eta^*) + O(N^{-1+2\delta})$. Of course, in these cases too, more exact results can be obtained by paying careful attention to the behavior of the extreme order statistics.

We conclude this section with a few applications of Theorems 4.1 and 4.2. The tedious computations will be omitted. First we consider the power $\pi_{W,N}(\theta)$ and $\pi_{W,L}(\theta)$ of Wilcoxon's signed rank test (W) against normal (N) and logistic (L) location alternatives $G(x) = \Phi(x-\theta)$ and $G(x) = (1 + \exp\{-(x-\theta)\})^{-1}$ respectively, where $\theta = O(N^{-\frac{1}{2}})$. We find

$$\begin{aligned} \pi_{W,N}(\theta) = 1 - \Phi(u_\alpha - \tilde\eta) - \frac{\tilde\eta\phi(u_\alpha - \tilde\eta)}{N} \Big\{ & \tfrac{26}{5} - 2^{\frac{1}{2}} - \tfrac{69}{20}u_\alpha^2 \\ & + \left(\tfrac{169}{20} - \frac{2(3)^{\frac{1}{2}}}{3}\right) u_\alpha \tilde\eta - \left(\tfrac{103}{20} - \frac{2(3)^{\frac{1}{2}}}{3} - \frac{\pi}{9}\right)\tilde\eta^2 \\ & + \frac{12 \operatorname{arctn} 2^{\frac{1}{2}}}{\pi}(-1 + u_\alpha^2 - 2u_\alpha\tilde\eta + \tilde\eta^2)\Big\} + o(N^{-1}), \end{aligned} \tag{4.21}$$

where $\tilde\eta = (3N/\pi)^{\frac{1}{2}}\theta$, and

$$\pi_{W,L}(\theta) = 1 - \Phi(u_\alpha - \eta^*) - \frac{\eta^*\phi(u_\alpha - \eta^*)}{20N}\{2 + 3u_\alpha^2 + u_\alpha\eta^* + \eta^{*2}\} + o(N^{-1}), \tag{4.22}$$

where $\eta^* = (N/3)^{\frac{1}{2}}\theta$.

As a second example we consider the one-sample normal scores test which is based on the scores $a_j = E\Phi^{-1}((1 + U_{j:N})/2)$. Its power $\pi_{NS,N}(\theta)$ and $\pi_{NS,L}(\theta)$ against the normal and logistic location alternatives described above satisfies

$$\begin{aligned} \pi_{NS,N}(\theta) = 1 - \Phi(u_\alpha - \eta^*) - \frac{\eta^*\phi(u_\alpha - \eta^*)}{4N} \Big\{ & -1 + u_\alpha^2 \\ & + 2\int_0^{\Phi^{-1}(1-1/2N)} \frac{(2\Phi(x) - 1)(1 - \Phi(x))}{\phi(x)}\,dx\Big\} + o(N^{-1}), \end{aligned} \tag{4.23}$$

where now $\eta^* = N^{\frac{1}{2}}\theta$, and

$$\begin{aligned}\pi_{NS,L}(\theta) = 1 - \Phi(u_\alpha - \tilde\eta) \\ - \frac{\tilde\eta\phi(u_\alpha - \tilde\eta)}{12N}\Big\{23 - 12(2)^{\frac{1}{2}} + u_\alpha^2 + (2\pi - 5)u_\alpha\tilde\eta \\ + (72 \arctan 2^{\frac{1}{2}} - 22\pi + 1)\tilde\eta^2 \\ - 6\int_0^{\Phi^{-1}(1-1/2N)} \frac{(2\Phi(x)-1)(1-\Phi(x))}{\phi(x)}\,dx\Big\} + o(N^{-1}),\end{aligned} \tag{4.24}$$

where now $\tilde\eta = (N/\pi)^{\frac{1}{2}}\theta$. We note that Theorem 4.2 ensures that (4.23) will also hold for van der Waerden's one-sample test which is based on the approximate scores $a_j = \Phi^{-1}((N+j+1)/2(N+1))$. To evaluate the integral in (4.23) and (4.24) we write

$$\begin{aligned}\int_0^{\Phi^{-1}(1-1/2N)} \frac{(2\Phi(x)-1)(1-\Phi(x))}{\phi(x)}\,dx \\ = \tfrac{1}{2}\log\log N + \tfrac{1}{2}\log 2 - 2\int_0^\infty \log x\,\phi(x)\,dx \\ + \int_0^\infty \frac{(2\Phi(x)-1)\{x(1-\Phi(x)) - \phi(x)\}}{x\phi(x)}\,dx + o(1) \\ = \tfrac{1}{2}\log\log N + \tfrac{1}{2}\log 2 + 0.05832\cdots + o(1),\end{aligned} \tag{4.25}$$

where the final result is obtained by numerical integration.

5. Permutation tests. In this section we consider distribution free tests other than rank tests, viz. permutation tests. We limit our discussion to linear permutation tests that reject the hypothesis of symmetry if

$$\sum_{i=1}^N h(X_i) \geqq \xi_\alpha(Z) \tag{5.1}$$

with possible randomization if equality occurs. Here h is a function on R^1, $Z = (Z_1, \cdots, Z_N)$ denotes the vector of order statistics of $|X_1|, \cdots, |X_N|$ as before and ξ_α is chosen in such a way that under the hypothesis of symmetry

$$P(\sum_{i=1}^N h(X_i) \geqq \xi_\alpha(Z) \mid Z) = \alpha \qquad \text{a.s.} \tag{5.2}$$

with an obvious modification if there is randomization.

Since (5.1) is equivalent to $\sum \{h(X_i) - h(-X_i)\} \geqq 2\xi_\alpha(Z) - \sum \{h(Z_j) + h(-Z_j)\}$, we assume without loss of generality that h is antisymmetric about the origin, i.e.

$$h(x) = -h(-x) \qquad \text{for all } x. \tag{5.3}$$

But then, under G and conditional on Z, $\sum h(X_i)$ is distributed as $2\sum a_j(V_j - \frac{1}{2})$ with V_j as in (2.3) and $a_j = h(Z_j)$. This means that we can obtain an expansion for this conditional distribution of $\sum h(X_i)$ if we can apply Theorem 2.1.

Under the hypothesis of symmetry, $P_j = \frac{1}{2}$ in (2.3) for all j. Hence in this case Theorem 2.1 yields an expansion for the conditional df of $\sum h(X_i)/(\sum h^2(Z_j))^{\frac{1}{2}}$ that holds uniformly on the set of all values of Z for which the $a_j = h(Z_j)$ satisfy

(2.35) and (2.36) for fixed c, C and δ. If α satisfies (3.6), this immediately leads to an expansion for $\xi_\alpha(Z)$. We find (cf. (3.13))

$$\text{(5.4)} \qquad \frac{\xi_\alpha(Z)}{(\sum h^2(Z_j))^{\frac{1}{2}}} = u_\alpha - \frac{\sum h^4(Z_j)}{12(\sum h^2(Z_j))^2}(u_\alpha^3 - 3u_\alpha) + O(N^{-\frac{5}{4}})$$

uniformly on the set E_0^c where, for fixed positive c, C and δ, $\sum h^2(Z_j) \geqq cN$, $\sum h^4(Z_j) \leqq CN$ and $\lambda\{x \mid \exists_j |x - h(Z_j)| < \zeta\} \geqq \delta N\zeta$ for some $\zeta \geqq N^{-\frac{3}{2}}\log N$.

Next we consider the contiguous location alternatives $G(x) = F(x - \theta)$ of Section 3. Under these alternatives, Theorem 2.1 yields an expansion for the conditional df of $\frac{1}{2}\{\sum h(X_i) - \sum (2P_j - 1)h(Z_j)\}/\{\sum P_j(1 - P_j)h^2(Z_j)\}^{\frac{1}{2}}$ uniformly on the set E_θ^c where, for fixed positive c, C and δ, $\sum P_j(1 - P_j)h^2(Z_j) \geqq cN$, $\sum h^4(Z_j) \leqq CN$ and $\lambda\{x \mid \exists_j |x - h(Z_j)| < \zeta, \varepsilon \leqq P_j \leqq 1 - \varepsilon\} \geqq \delta N\zeta$ for some $\zeta \geqq N^{-\frac{3}{2}}\log N$.

Since $E_0 \subset E_\theta$ it suffices to show that $P_\theta(E_\theta) = O(N^{-\frac{5}{4}})$ in order to obtain an expansion to $O(N^{-\frac{5}{4}})$ for the conditional power given Z of the permutation test. The unconditional power is then obtained by taking the expectation. This is done in very much the same way as in Sections 2 and 3 for linear rank tests, the only difference being that now not only the P_j but also the a_j depend on Z.

This program is carried out in Albers (1974) for the special case of the locally most powerful permutation test where $h = -\psi_1 = -f'/f$. In Theorem 5.1 we reproduce a version of this result without further proof. Of course a similar result may be obtained for the general linear permutation test (5.1) with $h \neq -\psi_1$.

Suppose that F is a df with a density f that is positive, symmetric about zero and five times differentiable. Define ψ_i and Ψ_i by (3.1) and (3.15) and take $h = -\psi_1$. Let $\pi_P(\theta)$ be the power of the permutation test (5.1) against the alternative $F(x - \theta)$ and define

$$\text{(5.5)} \qquad \begin{aligned} \pi_P^*(\theta) = {} & 1 - \Phi(u_\alpha - \eta^*) \\ & + \frac{\eta^*\phi(u_\alpha - \eta^*)}{72N}\left\{\frac{\int_0^1 \Psi_1^4(t)\,dt}{(\int_0^1 \Psi_1^2(t)\,dt)^2}[-6u_\alpha^2 - 3 + 3u_\alpha\eta^* + 5\eta^{*2}] \right. \\ & \left. - \frac{12\int_0^1 \Psi_2^2(t)\,dt}{(\int_0^1 \Psi_1^2(t)\,dt)^2}\eta^{*2} + 9(1 - u_\alpha\eta^* + \eta^{*2})\right\}, \end{aligned}$$

where η^* is given by (4.16).

THEOREM 5.1. *Let F satisfy (4.5) for $i = 1, \cdots, 5$ and $m_1 = 10$, $m_2 = \frac{5}{2}$, $m_3 = \frac{5}{3}$, $m_4 = \frac{5}{4}$, $m_5 = 1$ and suppose that positive numbers C and ε exist such that $0 \leqq \theta \leqq CN^{-\frac{1}{2}}$ and $\varepsilon \leqq \alpha \leqq 1 - \varepsilon$. Take $h = -\psi_1$. Then there exists $A > 0$ depending on N, F, θ and α only through F, C and ε and such that*

$$|\pi_P(\theta) - \pi_P^*(\theta)| \leqq AN^{-\frac{5}{4}}.$$

For $F = \Phi$, we have $-\psi_1(x) = x$ and Theorem 5.1 provides an expansion for the power of the permutation test based on $\sum X_i$ against normal shift alternatives

$\Phi(x - \theta)$ with $0 \leqq \theta \leqq CN^{-\frac{1}{2}}$ and $\varepsilon \leqq \alpha \leqq 1 - \varepsilon$. We find that this power equals

$$1 - \Phi(u_\alpha - N^{\frac{1}{2}}\theta) - \frac{\theta u_\alpha{}^2 \phi(u_\alpha - N^{\frac{1}{2}}\theta)}{4N^{\frac{1}{2}}} + O(N^{-\frac{5}{4}}) . \tag{5.6}$$

But (5.6) is also the power of Student's one-sided one-sample test for Φ against $\Phi(x - \theta)$ (cf. Hodges and Lehmann (1970)). It follows that for testing the hypothesis Φ against contiguous normal shift alternatives for fixed $0 < \alpha < 1$, the powers of the permutation test based on $\sum X_i$ and of Student's test differ by only $O(N^{-\frac{5}{4}})$ as $N \to \infty$. In fact, this difference is $O(N^{-\frac{3}{2}})$, since Φ satisfies the stronger regularity conditions needed to replace $N^{-\frac{5}{4}}$ by $N^{-\frac{3}{2}}$ in Theorem 5.1.

The remainder of this section will be devoted to a further investigation of this rather striking phenomenon. Roughly speaking, we shall show that for testing any given symmetric distribution against near alternatives, the permutation test (5.1) is almost equivalent to Student's test applied to $h(X_1), \cdots, h(X_N)$ with the correct level of significance for the given null-distribution. Our proof differs from the one outlined above in that we do not use power expansions to establish the near equivalence of the two tests. Instead, we show that the critical regions of the tests are almost identical. This more direct approach has the additional advantage of providing a simple explanation of our result.

Let F be the df of a distribution that is symmetric about zero and consider the problem of testing the hypothesis that $X_1, \cdots, X_N$ have df F against the alternative that they have another df G. For this testing problem and an arbitrary h satisfying (5.3) we compare the permutation test (5.1) with Student's test applied to $h(X_1), \cdots, h(X_N)$ that rejects the hypothesis if

$$\tilde{T} = \frac{\sum h(X_i)}{[\sum h^2(X_i) - N^{-1}(\sum h(X_i))^2]^{\frac{1}{2}}} (1 - N^{-1})^{\frac{1}{2}} \geqq t_\alpha \tag{5.7}$$

with possible randomization if equality occurs. Here t_α depends on α, h, F and N and is chosen in such a way that the test (5.7) has level α.

THEOREM 5.2. *Suppose there exist positive numbers $c, C, \varepsilon, \eta, \delta_1, \delta_2, \cdots$ with $\lim_{N\to\infty} \delta_N = 0$ and $m > 8$, such that hF^{-1} and hG^{-1} are monotone and differentiable on intervals I_F and I_G of length at least η where*

$$\left|\frac{d}{dt} h(F^{-1}(t))\right| \geqq c , \qquad \left|\frac{d}{dt} h(G^{-1}(t))\right| \geqq c , \tag{5.8}$$

and such that $\varepsilon \leqq \alpha \leqq 1 - \varepsilon$, and

$$\int_{-\infty}^{\infty} |h(x)|^m \, dF(x) \leqq C , \qquad \int_{-\infty}^{\infty} |h(x)|^m \, dG(x) \leqq C , \tag{5.9}$$

$$\left|\int_{-\infty}^{\infty} h^{2k}(x) \, dF(x) - \int_{-\infty}^{\infty} h^{2k}(x) \, dG(x)\right| \leqq \delta_N \qquad \text{for } k = 1, 2 . \tag{5.10}$$

Then there exist $A > 0$ depending on N, F, G, h and α only through c, C, η and ε, and $\beta > 0$ depending only on m, such that the powers of the tests (5.1) *and* (5.7) *for F against G differ by at most $A(N^{-\beta} + \delta_N)N^{-1}$.*

PROOF. We denote probabilities and expected values under $G(F)$ by $P_G(P_F)$ and $E_G(E_F)$. By (5.9) and (5.8) we have

$$\sigma_G^2(h(X_1)) \leqq E_G h^2(X_1) \leqq [E_G h^4(X_1)]^{\frac{1}{2}} \leqq C^{2/m}, \tag{5.11}$$

$$\sigma_G^2(h(X_1)) \geqq 2 \int_0^{\eta/2} (ct)^2\, dt = \frac{c^2\eta^3}{12}, \tag{5.12}$$

so that these moments are bounded away from 0 and ∞. For positive integer $k \leqq 4$, Markov's inequality, the Marcinkievitz–Zygmund–Chung inequality (Chung (1951)) and (5.9) yield

$$\begin{aligned} P_G(|\textstyle\sum (h^k(X_i) - E_G h^k(X_i))| \geqq \tau N) & \\ \leqq \frac{E_G|\sum (h^k(X_i) - E_G h^k(X_i))|^{m/k}}{(\tau N)^{m/k}} & \\ \leqq B_m(\tau^2 N)^{-m/(2k)} E_G|h^k(X_1) - E_G h^k(X_1)|^{m/k} & \\ \leqq B_m C\left(\frac{2}{\tau}\right)^{m/k} N^{-m/(2k)}, & \end{aligned} \tag{5.13}$$

where B_m depends only on m. Choose

$$\beta = \min\left(\frac{m-8}{2m+8}, \tfrac{1}{4}\right). \tag{5.14}$$

Taking $\tau = N^{-\beta}$ in (5.13) and using (5.3) we find that

$$\frac{1}{N} \sum h^{2k}(Z_j) = \frac{1}{N} \sum h^{2k}(X_i) = E_G h^{2k}(X_1) + O(N^{-\beta}), \qquad k = 1, 2, \tag{5.15}$$

$$\frac{1}{N} \sum h^2(X_i) - \left[\frac{1}{N} \sum h(X_i)\right]^2 = \sigma_G^2(h(X_1)) + O(N^{-\beta}), \tag{5.16}$$

uniformly on a set with probability $1 - O(N^{-1-\beta})$ under G.

Assumption (5.3) implies that

$$\lambda\{x \mid \exists_j |x - h(Z_j)| < \zeta\} \geqq \tfrac{1}{2}\lambda\{x \mid \exists_i |x - h(X_i)| < \zeta\},$$

and under G the right-hand side is distributed like

$$\tfrac{1}{2}\lambda\{x \mid \exists_j |x - h(G^{-1}(U_{j:N}))| < \zeta\},$$

where $U_{1:N} < \cdots < U_{N:N}$ are order statistics from a uniform distribution on (0, 1). Now for $n \geqq 1$

$$\begin{aligned} P(U_{j+n:N} - U_{j:N} \leqq z) & \\ = \iint_{0<s<t<1,\, t-s\leqq z} \frac{N!}{(j-1)!\,(n-1)!\,(N-j-n)!} s^{j-1}(t-s)^{n-1}(1-t)^{N-j-n}\, ds\, dt & \\ \leqq \frac{(Nz)^{n-1}}{(n-1)!} \iint_{0<s<t<1} \frac{(N-n+1)!}{(j-1)!\,(N-j-n)!} s^{j-1}(1-t)^{N-j-n}\, ds\, dt & \\ = \frac{(Nz)^{n-1}}{(n-1)!}. & \end{aligned}$$

Taking $n = 6$ and $z = 2c^{-1}N^{-\frac{5}{6}} \log N$ we see that

$$P\left(U_{6(k+1):N} - U_{6k:N} \geqq 2c^{-1}N^{-\frac{5}{6}} \log N \ \text{ for all } \ 1 \leqq k \leqq \left[\frac{N}{6}\right] - 1\right)$$
$$\geqq 1 - \frac{N}{6}(2c^{-1}N^{-\frac{5}{6}} \log N)^5 = 1 - O(N^{-1-\beta}) .$$

Together with (5.8) this implies that for $\zeta = N^{-\frac{5}{6}} \log N$

$$\lambda\{x \mid \exists_j |x - h(Z_j)| < \zeta\} \geqq \tfrac{1}{2}\eta N\zeta \tag{5.17}$$

with probability $1 - O(N^{-1-\beta})$ under G.

Now (5.11), (5.12), (5.15) and (5.17) ensure that expansion (5.4) holds uniformly except on a set E_0 with $P_G(E_0) = O(N^{-1-\beta})$. Simplifying this expansion by using (5.11), (5.12) and (5.15) once more, we arrive at the conclusion that the power against G of the test (5.1) is given by

$$\begin{aligned}\pi_P(G) = P_G\left(\frac{\sum h(X_i)}{(\sum h^2(X_i))^{\frac{1}{2}}} \geqq u_\alpha - \frac{E_G h^4(X_1)}{12N(E_G h^2(X_1))^2}(u_\alpha^{\,3} - 3u_\alpha) + O(N^{-1-\beta})\right)\\ + O(N^{-1-\beta}) .\end{aligned} \tag{5.18}$$

Here the first remainder term depends on Z but may now be taken to be uniformly $O(N^{-1-\beta})$.

The inequality $\sum h(X_i)/(\sum h^2(X_i))^{\frac{1}{2}} \geqq a$ is algebraically equivalent with

$$\frac{\sum h(X_i)}{[\sum h^2(X_i) - N^{-1}(\sum h(X_i))^2]^{\frac{1}{2}}} \geqq \frac{a}{(1 - a^2/N)^{\frac{1}{2}}}$$

on the set where $\sum h^2(X_i) - N^{-1}(\sum h(X_i))^2 \neq 0$ and provided that $a^2 < N$. We may apply this to (5.18) in view of the condition $\varepsilon \leqq \alpha \leqq 1 - \varepsilon$, (5.11), (5.12) and (5.16). At the same time we may replace E_G by E_F in (5.18), and by (5.10) this only involves adding $O(\delta_N N^{-1})$ to the first remainder term in (5.18). In this way we obtain

$$\begin{aligned}\pi_P(G) = P_G\Big(\tilde{T} \geqq u_\alpha + \frac{u_\alpha^{\,3} - u_\alpha}{2N} - \frac{E_F h^4(X_1)}{12N(E_F h^2(X_1))^2}(u_\alpha^{\,3} - 3u_\alpha)\\ + O\left(\frac{N^{-\beta} + \delta_N}{N}\right)\Big) + O(N^{-1-\beta}) ,\end{aligned} \tag{5.19}$$

where $\tilde{T}$ is the statistic in (5.7).

By (5.11), (5.12) and (5.16) we have for $B \geqq 0$,

$$\begin{aligned}&\sup_t P_G(t \leqq \tilde{T} \leqq t + BN^{-1}(N^{-\beta} + \delta_N))\\ &\leqq \sup_t P_G\left(t \leqq \frac{N^{-\frac{1}{2}} \sum h(X_i)}{\sigma_G(h(X_1))} \leqq t + 2BN^{-1}(N^{-\beta} + \delta_N)\right)\\ &\quad + O(N^{-1-\beta}) .\end{aligned} \tag{5.20}$$

Now (5.8) ensures that under G the distribution of $h(X_1)$ has an absolutely continuous part; in fact, this distribution may be written as a mixture $Q = \eta\tilde{Q}_1 + (1 - \eta)\tilde{Q}_2$ where $\tilde{Q}_1$ is an absolutely continuous distribution with density $\tilde{q}_1 \leqq (c\eta)^{-1}$. Moreover, (5.9) and Markov's inequality imply that $\tilde{Q}_1([-C_1, C_1]) \geqq \frac{1}{2}$

where $C_1 = \max(1, (2C/\eta)^{\frac{1}{3}})$. It follows that $Q = (\eta/2)Q_1 + (1 - \eta/2)Q_2$ where $Q_1([-C_1, C_1]) = 1$ and Q_1 is absolutely continuous with density $q_1 \leqq c_1 = 2(c\eta)^{-1}$.

Let ρ_1 be the ch.f. of Q_1. Obviously, for any fixed $t \neq 0$, $|\rho_1(t)| \leqq |\bar{\rho}_1(t)|$ where $\bar{\rho}_1$ is the ch.f. of the distribution with density

$$\begin{aligned} \bar{q}_1(y) &= c_1 \qquad \text{for} \quad y \in \bigcup_{k=0}^{n} \left[-C_1 + \frac{2k\pi}{|t|}, \, -C_1 + \frac{2k\pi + 2\xi}{|t|} \right] \\ &= 0 \qquad \text{elsewhere,} \end{aligned}$$

with $n = [C_1|t|/\pi]$ and $(n+1)c_1 2\xi/|t| = 1$. An easy calculation yields $|\bar{\rho}_1(t)| = (\sin \xi)/\xi$; for $|t| \geqq \pi/C_1$ we have $\xi \geqq \pi/(4c_1C_1)$. It follows that there exists $b > 0$ depending only on η, c and C, such that the ch.f. of $h(X_1)$ under G satisfies

$$(5.21) \qquad |E_G e^{ith(X_1)}| \leqq 1 - b \qquad \text{for } |t| \geqq \pi .$$

Because of (5.9), (5.12), (5.21) and Lemma 1 in Cramér (1962), page 27, the df of $\sigma_G^{-1}(h(X_1))N^{-\frac{1}{2}} \sum (h(X_i) - E_G h(X_i))$ under G has an Edgeworth expansion; uniformly for all G satisfying (5.8) and (5.9) for fixed c, C and η, the derivative of this expansion is bounded and its remainder term is $O(N^{-3})$. Applying this result and (5.20) to (5.19) we find

$$(5.22) \qquad \pi_P(G) = P_G(\tilde{T} \geqq \tilde{t}_\alpha) + O(N^{-1}(N^{-\beta} + \delta_N))$$

uniformly for fixed c, C, η and ε, where

$$(5.23) \qquad \tilde{t}_\alpha = u_\alpha + \frac{u_\alpha^3 - u_\alpha}{2N} - \frac{E_F h^4(X_1)}{12N(E_F h^2(X_1))^2}(u_\alpha^3 - 3u_\alpha) .$$

Let t_α be as defined in (5.7). Since F satisfies all assumptions imposed on G, (5.22) will hold under F as well as under G. We have $\pi_P(F) = \alpha$ and hence $\tilde{t}_\alpha = t_{\tilde{\alpha}}$ where $|\tilde{\alpha} - \alpha| = O(N^{-1}(N^{-\beta} + \delta_N))$ uniformly for $\varepsilon \leqq \alpha \leqq 1 - \varepsilon$, but of course also uniformly for $\varepsilon/2 \leqq \alpha \leqq 1 - \varepsilon/2$. Because t_α is decreasing in α and $\tilde{t}_\alpha$ has a bounded derivative with respect to α for $\varepsilon/2 \leqq \alpha \leqq 1 - \varepsilon/2$, it follows that

$$(5.24) \qquad t_\alpha = \tilde{t}_\alpha + O(N^{-1}(N^{-\beta} + \delta_N))$$

uniformly for $\varepsilon \leqq \alpha \leqq 1 - \varepsilon$. In view of (5.22) and the preceding part of the proof this implies that

$$(5.25) \qquad \pi_P(G) = P_G(\tilde{T} \geqq t_\alpha) + O(N^{-1}(N^{-\beta} + \delta_N))$$

uniformly for fixed c, C, η and ε. This completes the proof. □

It may be useful to comment briefly on assumption (5.10) in Theorem 5.2. Of course this assumption is satisfied for a sequence of alternatives G_N that tends to F in an appropriate manner. It is easy to see, for instance, that if the sequence G_N^N is contiguous to F^N, (5.9) implies (5.10) with $\delta_N = O(N^{-\frac{1}{2}})$. Similarly, (5.9) will imply (5.10) for some sequence $\delta_N = o(1)$ if h is continuous and G_N converges weakly to F.

6. Deficiencies of distribution free tests. Let F be a fixed df with density f that is positive, symmetric about zero and five times differentiable. Consider the problem of testing, on the basis of $X_1, \cdots, X_N$, the hypothesis $G = F$ against the alternative $G(x) = F(x - \theta)$ at level α. For any particular θ, the maximum power $\pi^+(\theta)$ is attained by the test based on the statistic $\sum \{\log f(X_i - \theta) - \log f(X_i)\}$. This statistic is a sum of i.i.d. random variables and therefore its df admits an Edgeworth expansion under the usual conditions. By expanding the cumulants of the statistic Albers (1974) obtains an expansion for $\pi^+(\theta)$. Define Ψ_i by (3.15) and take

$$\begin{aligned}\tilde{\pi}^+(\theta) = 1 - \Phi(u_\alpha - \eta^*) \\ &+ \frac{\eta^*\phi(u_\alpha - \eta^*)}{72N}\left\{\frac{\int_0^1 \Psi_1^4(t)\,dt}{(\int_0^1 \Psi_1^2(t)\,dt)^2}[3(u_\alpha^2 - 1) - 3\eta^* u_\alpha + 2\eta^{*2}]\right. \\ &\left. - \frac{3\int_0^1 \Psi_2^2(t)\,dt}{(\int_0^1 \Psi_1^2(t)\,dt)^2}\eta^{*2} - 9[(u_\alpha^2 - 1) - \eta^* u_\alpha]\right\},\end{aligned} \tag{6.1}$$

where η^* is given by (4.16). Lemma 6.1 is a version of Albers' result.

LEMMA 6.1. *Let F satisfy (4.5) for $m_i = 5/i$, $i = 1, \cdots, 5$, and suppose that positive numbers C and ε exist such that $0 \leqq \theta \leqq CN^{-\frac{1}{2}}$ and $\varepsilon \leqq \alpha \leqq 1 - \varepsilon$. Then there exists $A > 0$ depending on N, F, θ and α only through F, C and ε and such that*

$$|\pi^+(\theta) - \tilde{\pi}^+(\theta)| \leqq AN^{-\frac{3}{2}}. \tag{6.2}$$

For the same testing problem Theorem 4.2 provides an expansion for the power $\pi(\theta)$ of the locally most powerful rank test. Together, Theorem 4.2 and Lemma 6.1 will enable us to find the deficiency d_N of the locally most powerful rank test with respect to the most powerful parametric test. To ensure that F satisfies the assumptions of both Theorem 4.2 and Lemma 6.1, we require that $F \in \mathscr{F}_1$, where

DEFINITION 6.1. $\mathscr{F}_1$ is the class of df's F on R^1 with positive densities f that are symmetric about zero, five times differentiable and such that (4.5) is satisfied for $i = 1, \cdots, 5$ with $m_1 = 6$, $m_2 = 3$, $m_3 = \frac{5}{3}$, $m_4 = \frac{5}{4}$, $m_5 = 1$, and such that (4.6) holds.

Furthermore, define

$$\begin{aligned}\tilde{d}_N = \frac{1}{12}\left\{\frac{\int_0^1 \Psi_1^4(t)\,dt}{(\int_0^1 \Psi_1^2(t)\,dt)^2}[3(u_\alpha^2 - 1) - 2\eta^* u_\alpha - \eta^{*2}]\right. \\ &+ \frac{3\int_0^1 \Psi_2^2(t)\,dt}{(\int_0^1 \Psi_1^2(t)\,dt)^2}\eta^{*2} - 3[(u_\alpha^2 - 1) - 2\eta^* u_\alpha + \eta^{*2}] \\ &\left. + 12\,\frac{\int_{1/N}^{1-1/N} (\Psi_1'(t))^2 t(1-t)\,dt}{\int_0^1 \Psi_1^2(t)\,dt}\right\},\end{aligned} \tag{6.3}$$

with η^* as in (4.16).

THEOREM 6.1. *Let d_N be the deficiency of the locally most powerful rank test* *with respect to the most powerful parametric test for testing $G = F$ against $G(x) = F(x - \theta)$ on the basis of $X_1, \cdots, X_N$ and at level α. Suppose that $F \in \mathscr{F}_1$ and that $cN^{-\frac{1}{2}} \leqq \theta \leqq CN^{-\frac{1}{2}}$, $\varepsilon \leqq \alpha \leqq 1 - \varepsilon$ for positive c, C and ε. Then, for every fixed F, c, C and ε, there exist positive numbers A, $\delta_1, \delta_2, \cdots$ such that $\lim_{N\to\infty} \delta_N = 0$ and for every N*

$$|d_N - \tilde{d}_N| \leqq \delta_N + AN^{-\frac{1}{2}} \int_{1/N}^{1-1/N} (\Psi_1'(t))^2(t(1-t))^{\frac{1}{2}}\, dt\,. \tag{6.4}$$

This result continues to hold if the locally most powerful rank test is replaced by the rank test with the corresponding approximate scores $a_j = -\Psi_1(j/(N+1))$.

PROOF. As $\mathscr{F}_1 \subset \mathscr{F}$, the remark following Theorem 4.2 shows that

$$\int_{1/N}^{1-1/N} (\Psi_1'(t))^2(t(1-t))^{\nu}\, dt = o(N^{\frac{3}{2}-\nu}) \qquad \text{for } \nu = 1, \tfrac{1}{2}\,. \tag{6.5}$$

Theorem 4.2 and Lemma 6.1 provide expansions for $\pi(\theta)$ and $\pi^+(\theta)$. In view of (6.5), the boundedness of u_α and the fact that $c \leqq N^{\frac{1}{2}}\theta \leqq C$, it is clear from these expansions that $d_N = o(N^{\frac{1}{2}})$. To find d_N we replace N by $N + d_N$ and η^* by $\eta^*(1 + d_N N^{-1})^{\frac{1}{2}}$ in the expansion for $\pi(\theta)$ and equate the result to the expansion for $\pi^+(\theta)$. Taylor expansion with respect to $d_N N^{-1}$ in (4.18) yields

$$\begin{aligned} \frac{\eta^*\phi(u_\alpha - \eta^*)}{24N} \Bigg\{ & 12 d_N + \frac{\int_0^1 \Psi_1^4(t)\,dt}{(\int_0^1 \Psi_1^2(t)\,dt)^2}[-3(u_\alpha^2 - 1) + 2\eta^* u_\alpha + \eta^{*2}] \\ & - \frac{3\int_0^1 \Psi_2^2(t)\,dt}{(\int_0^1 \Psi_1^2(t)\,dt)^2}\eta^{*2} + 3(u_\alpha^2 - 1) - 6\eta^* u_\alpha + 3\eta^{*2} \\ & - \frac{12\int_{1/N}^{1-1/N}(\Psi_1'(t))^2 t(1-t)\,dt}{\int_0^1 \Psi_1^2(t)\,dt}\Bigg\} \\ & = o(N^{-1}) + O(N^{-\frac{3}{2}}\int_{1/N}^{1-1/N}(\Psi_1'(t))^2(t(1-t))^{\frac{1}{2}}\,dt)\,, \end{aligned} \tag{6.6}$$

uniformly for fixed $F \in \mathscr{F}_1$, c, C and ε. As $\eta^*\phi(u_\alpha - \eta^*)$ is bounded away from zero, (6.4) follows. The last assertion of the theorem is an immediate consequence of Theorem 4.2. □

Obviously (6.3) and (6.4) imply that under the conditions of Theorem 6.1

$$d_N = O(\int_{1/N}^{1-1/N}(\Psi_1'(t))^2 t(1-t)\,dt) \tag{6.7}$$

for $N \to \infty$. Hence d_N remains bounded as $N \to \infty$ if $\int_0^1 (\Psi_1'(t))^2 t(1-t)\,dt$ converges. Fortunately, in most cases of interest Theorem 6.1 provides more detailed information than (6.7) and remarks similar to those following Theorem 4.2 apply. Typically Ψ_1' will be bounded near 0 and the asymptotic behavior of d_N will be determined by the rate of growth of Ψ_1' near 1. If $\Psi_1'(t) = o((1-t)^{-1})$ near 1, then $d_N = \tilde{d}_N + o(1)$. If $\Psi_1'(t)$ is of exact order $(1-t)^{-1}$, then

$$d_N = \frac{\int_0^{1-1/N}(\Psi_1'(t))^2 t(1-t)\,dt}{\int_0^1 \Psi_1^2(t)\,dt} + O(1)$$

and d_N will be of the order $\log N$. Finally, if $\Psi_1'(t) \sim (1-t)^{-1-\delta}$ for $t \to 1$ and some $0 < \delta < \frac{1}{6}$, then the expansion (6.4) reduces to $d_N = O(N^{2\delta})$, which is nothing but (6.7).

We shall give two applications of Theorem 6.1. First we consider the problem of testing the hypothesis $G = \Phi$ against the alternative $G(x) = \Phi(x - \theta)$, where $cN^{-\frac{1}{2}} \leq \theta \leq CN^{-\frac{1}{2}}$. Let d_N be the deficiency of the normal scores test (or van der Waerden's test) with respect to the most powerful parametric test based on $\bar{X}$. Computations similar to those in Section 4 yield

$$(6.8) \qquad d_N = \tfrac{1}{2}(u_\alpha^2 - 1) + \int_0^{\Phi^{-1}(1-1/2N)} \frac{(2\Phi(x) - 1)(1 - \Phi(x))}{\phi(x)}\, dx + o(1)$$
$$= \tfrac{1}{2} \log \log N + \tfrac{1}{2}(u_\alpha^2 - 1) + \tfrac{1}{2} \log 2 + 0.05832 \cdots + o(1)\,.$$

In this case $d_N \sim \frac{1}{2} \log \log N \to \infty$ for $N \to \infty$. Note that there is no dependence on θ in this expansion for d_N and that the leading term is also independent of α.

As a second example we take the logistic df $F(x) = (1 + e^{-x})^{-1}$ and consider the testing problem $G = F$ against $G(x) = F(x - bN^{-\frac{1}{2}})$, where $b > 0$ is fixed. Now d_N is the deficiency of Wilcoxon's signed rank test with respect to the most powerful parametric test for this problem. We find

$$(6.9) \qquad d_N = \tfrac{1}{60}\{18 + 12u_\alpha^2 + 4(3)^{\frac{1}{2}} b u_\alpha + b^2\} + o(1)$$

and here d_N tends to a finite limit for $N \to \infty$.

Having shown that the deficiency of a distribution free test with respect to the best parametric test may tend to a finite limit, we now address ourselves to the intriguing question whether this limit can be zero. To answer this question we first have to decide what is meant by the best parametric test. So far, we have compared the performance of a distribution free test with that of the most powerful parametric test for known scale against a simple location alternative, thus in effect comparing with envelope power. Of course this comparison is not quite fair. Computed in this way, the deficiency of a distribution free test reflects the losses incurred by using (i) the same test against every location alternative $\theta > 0$; (ii) a scale invariant test; (iii) a distribution free test. Since our main interest is the deficiency due to (iii), it is more appropriate to compare with the uniformly most powerful scale invariant test, if such a test exists. Unfortunately, invariant tests are in general rather intractable, the main exception being Student's test for the normal location case. We note that Hodges and Lehmann (1970) have shown that the deficiency of Student's test with respect to the most powerful parametric test based on $\bar{X}$ tends to a finite but positive limit, so that it does indeed matter whether one compares with Student's test or with envelope power.

We are thus led to consider the normal location case with Student's test as the best parametric test. To establish the existence of a distribution free test with deficiency tending to zero, the obvious candidate is the permutation test based on $\sum X_i$. Theorem 6.2 is an immediate consequence of Theorem 5.1 and the remark following it.

THEOREM 6.2. *Let d_N be the deficiency of the permutation test based on $\sum X_i$ with respect to Student's test for testing $G = \Phi$ against $G(x) = \Phi(x - \theta)$ on the* *basis of $X_1, \cdots, X_N$ and at level α. Suppose that positive numbers c, C and ε exist such that $cN^{-\frac{1}{2}} \leqq \theta \leqq CN^{-\frac{1}{2}}$ and $\varepsilon \leqq \alpha \leqq 1 - \varepsilon$. Then there exists $A > 0$ depending on N, θ and α only through c, C and ε and such that*

$$d_N \leqq AN^{-\frac{1}{4}} . \tag{6.10}$$

Hence in this case we do find that d_N tends to zero for $N \to \infty$. Perhaps the most surprising thing about this example is that asymptotically one has to pay a certain price for scale invariance, but that once this price has been paid, there is no additional penalty for using a distribution free test. We note that the remark following Theorem 5.1 implies that (6.10) may be replaced by $d_N \leqq AN^{-\frac{1}{2}}$.

Theorem 6.2 may of course be generalized considerably by taking Theorem 5.2 for $h(x) \equiv x$ as a starting point instead of Theorem 5.1. For d_N as in Theorem 6.2, it is clear that $d_N = o(1)$ for a much larger class of testing problems than the normal location problem of Theorem 6.2. Although Student's test is generally not optimal for these problems, this shows how closely the two tests resemble one another.

7. Expansions and deficiencies for related estimators. Let $T = T(X_1, \cdots, X_N)$ be given by (2.2) and suppose that the scores a_j are nonnegative and nondecreasing in $j = 1, \cdots, N$. Define the statistic M by

$$\begin{aligned} M(X_1, \cdots, X_N) = \tfrac{1}{2} \sup \{t : 2T(X_1 - t, \cdots, X_N - t) > \textstyle\sum a_j\} \\ + \tfrac{1}{2} \inf \{t : 2T(X_1 - t, \cdots, X_N - t) < \textstyle\sum a_j\} . \end{aligned} \tag{7.1}$$

Suppose that $X_1, \cdots, X_N$ are i.i.d. with common df $G(x) = F(x - \mu)$, where F has a density f that is symmetric about zero. Then M is the midpoint of the interval between the upper and lower 0.5 confidence bounds for μ induced by the statistic T. Hodges and Lehmann (1963) proposed M as an estimator for μ and studied its connection with T. They showed that the normal approximation to the power of the level $\frac{1}{2}$ test based on T for contiguous location alternatives could be used to establish asymptotic normality of M. We shall show that, similarly, power expansions for level $\frac{1}{2}$ yield expansions for the df of $N^{\frac{1}{2}}(M - \mu)$. We restrict attention to the case where the scores are generated by a smooth function J.

Let $\mathscr{J}$ and $\mathscr{F}$ be given by Definition 4.1, let $\pi(\theta, \frac{1}{2})$ denote the power of the level $\frac{1}{2}$ right-sided test based on T against the alternative $F(x - \theta)$ and define $K_{\theta,i}$ and $\tilde{\eta}$ as in (4.8)—(4.10).

THEOREM 7.1. *Let $F \in \mathscr{F}$, $J \in \mathscr{J}$, suppose that J is nonnegative and nondecreasing and let $a_j = EJ(U_{j:N})$. Take $\theta = \xi N^{-\frac{1}{2}}$. Then, for every fixed J, F and $C > 0$,*

$$\sup_{|\xi| \leqq C} |P_\mu(N^{\frac{1}{2}}(M - \mu) \leqq \xi) - \pi(\theta, \tfrac{1}{2})| = O(N^{-\frac{5}{4}}) , \tag{7.2}$$

$$\sup_{|\xi| \leqq C} |P_\mu(N^{\frac{1}{2}}(M - \mu) \leqq \xi) - \{1 - K_{\theta,1}(-\tilde{\eta})\}| = o(N^{-1}) , \tag{7.3}$$

$$\begin{aligned} \sup_{|\xi| \leqq C} |P_\mu(N^{\frac{1}{2}}(M - \mu) \leqq \xi) - \{1 - K_{\theta,2}(-\tilde{\eta})\}| \\ = o(N^{-1}) + O(N^{-\frac{3}{2}} \textstyle\int_{1/N}^{1-1/N} |J'(t)|(|J'(t)| + |\Psi_1'(t)|)(t(1 - t))^{\frac{1}{2}}\, dt) . \end{aligned} \tag{7.4}$$

PROOF. It follows from Hodges and Lehmann (1963) that M is translation invariant and that its distribution is absolutely continuous and symmetric about μ. Thus, for $\theta = \xi N^{-\frac{1}{2}}$,

$$P_\mu(N^{\frac{1}{2}}(M - \mu) \leqq \xi) = P_\theta(M \geqq 0)\,, \tag{7.5}$$

and, in view of (7.1),

$$P_\theta(2T > \textstyle\sum a_j) \leqq P_\theta(M \geqq 0) \leqq P_\theta(2T \geqq \sum a_j)\,. \tag{7.6}$$

According to the proof of Theorem 4.1, the conclusions of Theorems 3.1 and 3.2 hold, which implies that $P_\theta(2T = \sum a_j) = O(N^{-\frac{5}{4}})$ uniformly for $|\theta| \leqq CN^{-\frac{1}{2}}$. This proves (7.2). The remaining part of Theorem 7.1 is now an immediate consequence of Theorem 4.1. □

The case where $J = -\Psi_1$, with Ψ_1 as in (3.15), is of course of special interest. Theorem 7.2 deals with this case for exact as well as approximate scores. Note that for $F \in \mathscr{F}$, the condition that $-\Psi_1$ is nonnegative and nondecreasing is equivalent to concavity of $\log f$, i.e. to strong unimodality of f.

THEOREM 7.2. *Let $F \in \mathscr{F}$, suppose that f is strongly unimodal and let either $a_j = -E\Psi_1(U_{j:N})$ for $j = 1, \cdots, N$ or $a_j = -\Psi_1(j/(N+1))$ for $j = 1, \cdots, N$. Then, for every fixed F and $C > 0$,*

$$\sup_{|\xi| \leqq C} |P_\mu(N^{\frac{1}{2}}(M - \mu) \leqq \xi) - \pi(\xi N^{-\frac{1}{2}}, \tfrac{1}{2})| = O(N^{-\frac{5}{4}})\,, \tag{7.7}$$

$$\begin{aligned} &P_\mu((N \textstyle\int_0^1 \Psi_1^2(t)\,dt)^{\frac{1}{2}}(M - \mu) \leqq x) \\ &\quad = \Phi(x) + \frac{x\phi(x)}{72N}\Big\{x^2\Big[\frac{5\int_0^1 \Psi_1^4(t)\,dt}{(\int_0^1 \Psi_1^2(t)\,dt)^2} - \frac{12\int_0^1 \Psi_2^2(t)\,dt}{(\int_0^1 \Psi_1^2(t)\,dt)^2} + 9\Big] \\ &\qquad + \frac{6\int_0^1 \Psi_1^4(t)\,dt}{(\int_0^1 \Psi_1^2(t)\,dt)^2} - \frac{36\int_{1/N}^{1-1/N} (\Psi_1'(t))^2 t(1-t)\,dt}{\int_0^1 \Psi_1^2(t)\,dt}\Big\} \\ &\qquad + o(N^{-1}) + O(N^{-\frac{3}{2}} \textstyle\int_{1/N}^{1-1/N} (\Psi_1'(t))^2(t(1-t))^{\frac{1}{2}}\,dt) \end{aligned} \tag{7.8}$$

uniformly for $|x| \leqq C$.

PROOF. The proof of (7.7) is identical to the proof of (7.2) in Theorem 7.1. Expansion (7.8) follows from (7.7) and Theorem 4.2. □

The estimators in Theorem 7.2 are efficient and their natural competitor is the maximum likelihood estimator M' which solves

$$\textstyle\sum_{j=1}^N \psi_1(X_j - M') = 0 \tag{7.9}$$

with ψ_1 as in (3.1). The performance of M' is connected with that of the locally most powerful test for F against $F(x - \theta)$, which is based on the statistic $-\sum \psi_1(X_j)$. Let $\pi'(\theta, \frac{1}{2})$ be the power of the level $\frac{1}{2}$ right-sided test based on $-\sum \psi_1(X_j)$ for F against $F(x - \theta)$.

LEMMA 7.1. *Suppose that f is positive, symmetric about zero and strongly unimodal and that* (4.5) *is satisfied for $m_i = 5/i$, $i = 1, \cdots, 5$. Then, for every fixed F and*

$C > 0$,

$$\sup_{|\xi| \leqq C} |P_\mu(N^{\frac{1}{2}}(M' - \mu) \leqq \xi) - \pi'(\xi N^{-\frac{1}{2}}, \tfrac{1}{2})| = O(N^{-\frac{5}{4}}), \tag{7.10}$$

$$\begin{aligned} &P_\mu((N \textstyle\int_0^1 \Psi_1^2(t)\,dt)^{\frac{1}{2}}(M' - \mu) \leqq x) \\ &\quad = \Phi(x) + \frac{x\phi(x)}{72N}\left\{x^2\left[\frac{5 \int_0^1 \Psi_1^4(t)\,dt}{(\int_0^1 \Psi_1^2(t)\,dt)^2} - \frac{12 \int_0^1 \Psi_2^2(t)\,dt}{(\int_0^1 \Psi_1^2(t)\,dt)^2} + 9\right]\right. \\ &\qquad \left. - \frac{3 \int_0^1 \Psi_1^4(t)\,dt}{(\int_0^1 \Psi_1^2(t)\,dt)^2} + 9\right\} + O(N^{-\frac{5}{4}}) \end{aligned} \tag{7.11}$$

uniformly for $|x| \leqq C$.

Proof. The estimator M' is translation invariant and its distribution is symmetric about μ. Thus, for $\theta = \xi N^{-\frac{1}{2}}$, (7.5) holds with M replaced by M', and in view of (7.9),

$$P_\theta(-\textstyle\sum \psi_1(X_j) > 0) \leqq P_\mu(N^{\frac{1}{2}}(M' - \mu) \leqq \xi) \leqq P_\theta(-\sum \psi_1(X_j) \geqq 0). \tag{7.12}$$

Since f is everywhere positive and ψ_1 is everywhere differentiable, the distribution of $\psi_1(X_1)$ under θ contains a fixed absolutely continuous component for all θ in a neighborhood of zero. Together with (4.5) for $m_1 = 5$, this ensures that the df of $\sum \psi_1(X_j)$ under θ possesses an Edgeworth expansion with remainder $O(N^{-\frac{5}{4}})$ uniformly for $|\theta| \leqq CN^{-\frac{1}{2}}$. This implies that $P_\theta(-\sum \psi_1(X_j) = 0) = O(N^{-\frac{5}{4}})$ uniformly for $|\theta| \leqq CN^{-\frac{1}{2}}$, which proves (7.10).

The expansion for the df of $\sum \psi_1(X_j)$ is used in Albers (1974) to establish an expansion for the power of the locally most powerful test under the conditions of Lemma 6.1. Specializing to the case where $\alpha = \frac{1}{2}$ and using (7.10) we obtain (7.11). □

There is no unique natural measure of scale to assess the performance of an estimator $\hat{\mu}$ admitting an expansion of the form (7.8) or (7.11). One possibility is to consider a family of measures determined by the quantiles of $\hat{\mu}$. We can define $\sigma(\hat{\mu}, s)$ to be the s-quantile of $(\hat{\mu} - \mu)$ divided by $u_{1-s} = \Phi^{-1}(s)$. As we are only considering estimators that are distributed symmetrically about μ, $\sigma(\hat{\mu}, s)$ may serve as a measure of scale for any $\frac{1}{2} < s < 1$. If we fix a value of s, we can define the deficiency $D_N(s)$ of a sequence of estimators $\{\hat{\mu}_{2,N}\}$ with respect to an estimator $\hat{\mu}_{1,N}$ by equating $\sigma(\hat{\mu}_{2,N+D_N}, s)$ and $\sigma(\hat{\mu}_{1,N}, s)$, with the usual convention that σ is determined by linear interpolation for nonintegral values of $N + D_N$. Similarly, for two sequences of level α tests, $d_N(\alpha, s)$ will denote the deficiency as defined in Section 1 for the case where the alternative θ is chosen in such a way that the common power equals s.

Let $\mathscr{F}_1$ be given by Definition 6.1.

Theorem 7.3. *Let $d_N(\frac{1}{2}, s)$ be the deficiency for level $\frac{1}{2}$ and power s of the locally most powerful rank test with respect to the locally most powerful test for testing F against $F(x - \theta)$. Let $D_N(s)$ be the deficiency of the Hodges–Lehmann estimator associated with the locally most powerful rank test with respect to the maximum likelihood estimator for estimating μ in $F(x - \mu)$. Suppose that $F \in \mathscr{F}_1$ and that f is* *strongly unimodal. Then, for fixed F and* $\frac{1}{2} < s < 1$,

$$|D_N(s) - d_N(\tfrac{1}{2}, s)| = O(N^{-\frac{1}{4}}), \tag{7.13}$$

$$D_N(s) = \frac{\int_{1/N}^{1-1/N} (\Psi_1'(t))^2 t(1-t)\,dt}{\int_0^1 \Psi_1^2(t)\,dt} - \frac{1}{4}\,\frac{\int_0^1 \Psi_1^4(t)\,dt}{(\int_0^1 \Psi_1^2(t)\,dt)^2} + \tfrac{1}{4} + o(1) + O(N^{-\frac{1}{2}} \int_{1/N}^{1-1/N} (\Psi_1'(t))^2 (t(1-t))^{\frac{1}{2}}\,dt). \tag{7.14}$$

This result continues to hold if in the locally most powerful rank test and the associated estimator, the exact scores are replaced by the approximate scores $a_j = -\Psi_1(j/(N+1))$.

PROOF. The conditions of Theorem 7.2 and Lemma 7.1 are satisfied. Writing M_N and M_N' for M and M', we see that for some ξ

$$P_\mu(N^{\frac{1}{2}}(M_N' - \mu) \leqq \xi) = s + O(N^{-\frac{3}{4}}), \tag{7.15}$$

$$P_\mu(N^{\frac{1}{2}}(M_{N+d_N} - \mu) \leqq \xi) = s + O(N^{-\frac{5}{4}}). \tag{7.16}$$

By the remark following Theorem 4.2 we have $\Psi_1'(t) = o((t(1-t))^{-\frac{3}{2}})$ near 0 and 1, and combining this with (7.8) and (7.11) we find that (7.15) and (7.16) imply (7.13). The proof of (7.14) is now the same as that of Theorem 6.1. □

An interesting property of the expansion (7.14) is that it is independent of s. Thus, to the order considered, the deficiency $D_N(s)$ is asymptotically independent of the particular choice of the quantile used to measure the performance of the estimators. Of course, this reflects the fact that the deficiency $d_N(\frac{1}{2}, s)$ is independent of the power in the same asymptotic sense. Algebraically, the reason for this phenomenon is that the term involving $x^3\phi(x)$ is the same in (7.8) and (7.11).

We also note that upon formal substitution of $\alpha = \frac{1}{2}$ and $\theta = 0$ in (6.3), the expansion for d_N in Theorem 6.1 reduces to the expansion for $D_N(s)$ in Theorem 7.3. This shows that if the remainder in (7.14) is $o(1)$, then $D_N(s)$ will tend to a nonnegative but possibly infinite limit.

In Section 6 we have already pointed out that an expansion like (7.14) may or may not be of interest, depending on the behavior of the remainder term. We should stress that, even if the expansion (7.14) is useless, (7.13) still establishes the asymptotic equivalence of $D_N(s)$ and $d_N(\frac{1}{2}, s)$.

We conclude our discussion with one example of Theorem 7.3. For estimating normal location, the deficiency of either one of the Hodges–Lehmann estimators associated with the normal scores test and with van der Waerden's test with respect to $\bar{X}$ is asymptotic to $\frac{1}{2}\log\log N$. The deficiency of one of these Hodges–Lehmann estimators with respect to the other tends to zero for $N \to \infty$.

APPENDIX

1. Expansions for the contiguous case. Our purpose in this appendix will be the justification of the passage from (2.41) to (3.8) under the assumptions stated in Section 3. Thus we shall suppose throughout that f is positive and symmetric about 0 and that $g(x) = f(x - \theta)$.

Begin by defining a function $\xi(x, t)$ for $x \geqq 0$, $t \geqq 0$, by

$$F(\xi(x, t) - t) + F(\xi(x, t) + t) = 2F(x) . \tag{A1.1}$$

Introduce also two other functions of two variables, p and $\tilde{p}$, by

$$p(x, t) = \frac{f(x - t)}{f(x - t) + f(x + t)} , \tag{A1.2}$$

$$\tilde{p}(x, t) = p(\xi(x, t), t) . \tag{A1.3}$$

The basic property of the function ξ is, of course, that the joint distribution of $(\xi(Z_1, \theta), \cdots, \xi(Z_N, \theta))$ under F is the same as the joint distribution of $(Z_1, \cdots, Z_N)$ under G. It follows that the joint distribution of $(\tilde{p}(Z_1, \theta), \cdots, \tilde{p}(Z_N, \theta))$ under F is the same as the joint distribution of $(P_1, \cdots, P_N)$ under G. It is evident therefore that our task is essentially that of expanding $\tilde{p}(x, t)$ around 0 as a function of t and giving suitable estimates of the remainder terms. We begin by differentiating formally. For convenience we shall, for any function of two variables $q(x, t)$, write

$$q_{i,j}(x, t) = \frac{\partial^{i+j} q(x, t)}{\partial x^i \, \partial t^j} .$$

Differentiating (A1.1) with respect to t we get

$$\xi_{0,1} = 2\tilde{p} - 1 . \tag{A1.4}$$

It is now easy though tedious to obtain $\tilde{p}_{0,j}(x, t)$ in terms of the $p_{i,k}(\xi(x, t), t)$ by replacing $\xi_{0,1}$ by $2\tilde{p} - 1$ after each differentiation. Thus, for example,

$$\tilde{p}_{0,1}(x, t) = [p_{0,1} + p_{1,0}(2p - 1)](\xi(x, t), t) , \tag{A1.5}$$

$$\begin{aligned} \tilde{p}_{0,2}(x, t) = [&p_{0,2} + 2p_{1,1}(2p - 1) + p_{2,0}(2p - 1)^2 + 2p_{1,0}p_{0,1} \\ &+ 2p_{1,0}^2(2p - 1)](\xi(x, t), t) . \end{aligned} \tag{A1.6}$$

Calculation of the $p_{i,j}$ is also tedious. Again we list the first few. Define

$$_1\psi_k(x, t) = \psi_k(x - t) , \qquad _2\psi_k(x, t) = \psi_k(x + t) , \tag{A1.7}$$

where $\psi_k = f^{(k)}/f$ as defined in (3.1), and let

$$_1\tilde{\psi}_k(x, t) = \psi_k(\xi(x, t) - t) , \qquad _2\tilde{\psi}_k(x, t) = \psi_k(\xi(x, t) + t) . \tag{A1.8}$$

Then

$$\begin{aligned} p_{0,1} &= -p(1 - p)[_1\psi_1 + {_2\psi_1}] , \qquad p_{1,0} = p(1 - p)[_1\psi_1 - {_2\psi_1}] , \\ p_{0,2} &= p(1 - p)[_1\psi_2 - {_2\psi_2} - 2p \cdot {_1\psi_1^2} + 2(1 - p){_2\psi_1^2} \\ &\quad + 2(1 - 2p){_1\psi_1} \cdot {_2\psi_1}] , \end{aligned} \tag{A1.9}$$

$$\begin{aligned} p_{1,1} &= p(1 - p)[-{_1\psi_2} - {_2\psi_2} + 2p \cdot {_1\psi_1^2} + 2(1 - p){_2\psi_1^2}] , \\ p_{2,0} &= p(1 - p)[_1\psi_2 - {_2\psi_2} - 2p \cdot {_1\psi_1^2} + 2(1 - p){_2\psi_1^2} \\ &\quad - 2(1 - 2p){_1\psi_1} \cdot {_2\psi_1}] . \end{aligned} \tag{A1.10}$$

Substituting (A1.9) and (A1.10) into (A1.5) and (A1.6) at $t = 0$ and employing similar manipulations with the third order derivatives we obtain

$$\text{(A1.11)} \qquad \tilde{p}(x, 0) = \tfrac{1}{2}, \qquad \tilde{p}_{0,1}(x, 0) = -\tfrac{1}{2}\psi_1(x), \qquad \tilde{p}_{0,2}(x, 0) = 0,$$
$$\tilde{p}_{0,3}(x, 0) = -\tfrac{1}{2}\psi_3(x) + 3\psi_1(x)\psi_2(x) - \tfrac{3}{2}\psi_1^3(x).$$

Moreover, from (A1.9), (A1.10) and the boundedness of p it is easy to see that constants b_1 and b_2 exist such that

$$\text{(A1.12)} \qquad |\tilde{p}_{0,1}| \leqq b_1 \textstyle\sum_{i=1}^2 |{}_i\tilde{\psi}_1|, \qquad |\tilde{p}_{0,2}| \leqq b_2 \textstyle\sum_{i=1}^2 \{|{}_i\tilde{\psi}_2| + {}_i\tilde{\psi}_1^2\}.$$

Similarly bounding first the $p_{i,k}$ and expressing $\tilde{p}_{0,j}$ appropriately, and invoking the inequality $|ab| \leqq r^{-1}|a|^r + s^{-1}|b|^s$, $r^{-1} + s^{-1} = 1$, we obtain for suitable b_3 and b_4

$$\text{(A1.13)} \qquad |\tilde{p}_{0,3}| \leqq b_3 \textstyle\sum_{i=1}^2 \{|{}_i\tilde{\psi}_3| + |{}_i\tilde{\psi}_2|^{\frac{3}{2}} + |{}_i\tilde{\psi}_1|^3\},$$
$$|\tilde{p}_{0,4}| \leqq b_4 \textstyle\sum_{i=1}^2 \{|{}_i\tilde{\psi}_4| + |{}_i\tilde{\psi}_3|^{\frac{4}{3}} + {}_i\tilde{\psi}_2^2 + {}_i\tilde{\psi}_1^4\}.$$

We need the following application of Taylor's formula with Cauchy's form of the remainder.

LEMMA A1.1. *Let $q(x, t)$ be a function of two variables possessing derivatives of order $\leqq k + 1$ in t in a neighborhood of* 0. *Then if S is any* rv *and $m \geqq 1$,*

$$\text{(A1.14)} \qquad E\left|q(S, t) - \textstyle\sum_{j=0}^k q_{0,j}(S, 0)\dfrac{t^j}{j!}\right|^m$$
$$\leqq \left[\frac{|t|^{k+1}}{(k+1)!}\right]^m \sup\{E|q_{0,k+1}(S, \nu t)|^m : 0 \leqq \nu \leqq 1\}.$$

Suppose moreover that for $j = 0, \cdots, k$, $Eq_{0,j}(S, 0)$ exists and is finite. Then

$$\text{(A1.15)} \qquad E\left|\{q(S, t) - Eq(S, t)\} - \textstyle\sum_{j=0}^k \{q_{0,j}(S, 0) - Eq_{0,j}(S, 0)\}\dfrac{t^j}{j!}\right|^m$$
$$\leqq 2^m \left[\frac{|t|^{k+1}}{(k+1)!}\right]^m \sup\{E|q_{0,k+1}(S, \nu t)|^m : 0 \leqq \nu \leqq 1\}.$$

PROOF. We have (cf. Dieudonné (1960), page 186, Titchmarsh (1939), page 368)

$$\text{(A1.16)} \qquad q(S, t) = \textstyle\sum_{j=0}^k q_{0,j}(S, 0)\dfrac{t^j}{j!}$$
$$+ \frac{t^{k+1}}{(k+1)!} \int_0^1 (k+1)(1-\nu)^k q_{0,k+1}(S, \nu t)\, d\nu$$

provided that the integral converges. Hence the left-hand side of (A1.14) is bounded by

$$\left[\frac{|t|^{k+1}}{(k+1)!}\right]^m E\left|\int_0^1 (k+1)(1-\nu)^k q_{0,k+1}(S, \nu t)\, d\nu\right|^m.$$

This obviously remains true even if the integral diverges for some values of S. An application of Ljapunov's inequality and Fubini's theorem complete the proof of (A1.14) and a similar argument disposes of (A1.15). □

Note that by using the same device one can show that the left-hand side of (A1.14) and (A1.15) is $o(|t|^{mk})$ for $t \to 0$ if q is k times continuously differentiable and

$$\lim_{t \to 0} E|q_{0,k}(S, t)|^m = E|q_{0,k}(S, 0)|^m . \tag{A1.17}$$

Of course (A1.17) holds if $q_{0,k}(S, \cdot)$ is continuous at 0 and

$$\sup \{E|q_{0,k}(S, t)|^{m+\delta} : |t| \leqq \delta\} < \infty \tag{A1.18}$$

for some $\delta > 0$.

We introduce two final pieces of notation. If $d_1, \cdots, d_N$ is a sequence of numbers we write

$$||d|| = \frac{1}{N} \sum_{j=1}^N |d_j| . \tag{A1.19}$$

If χ is a function of one variable and $\varepsilon > 0$ is fixed we define

$$||\chi|| = \sup \{\textstyle\int_{-\infty}^{\infty} |\chi(x + y)| f(x)\, dx : |y| \leqq \varepsilon\} . \tag{A1.20}$$

THEOREM A1.1. *Suppose that f is four times differentiable, that $E_0\psi_3(|X_1|)$, $E_0\psi_1(|X_1|)\psi_2(|X_1|)$ and $E_0\psi_1^3(|X_1|)$ exist and are finite and that $0 \leqq 2\theta \leqq \varepsilon$. Then if $r \geqq 1$, $r^{-1} + s^{-1} = 1$, there exists a constant B such that*

$$\begin{aligned} \sum_{j=1}^N a_j(2\pi_j - 1) = -\theta \sum_{j=1}^N a_j E_0\psi_1(Z_j) - \frac{\theta^3}{6} \sum_{j=1}^N a_j E_0[\psi_3(Z_j) \\ - 6\psi_1(Z_j)\psi_2(Z_j) + 3\psi_1^3(Z_j)] + M_1 , \\ |M_1| \leqq BN\theta^4 ||a^r||^{1/r}[||\psi_4^s|| + ||\psi_3^{4s/3}|| + ||\psi_2^{2s}|| + ||\psi_1^{4s}||]^{1/s} ; \end{aligned} \tag{A1.21}$$

$$\begin{aligned} \sum_{j=1}^N a_j^3(2\pi_j - 1) = -\theta \sum_{j=1}^N a_j^3 E_0\psi_1(Z_j) + M_2 , \\ |M_2| \leqq BN\theta^3 ||a^{3r}||^{1/r}[||\psi_3^s|| + ||\psi_2^{3s/2}|| + ||\psi_1^{3s}||]^{1/s} ; \end{aligned} \tag{A1.22}$$

$$\begin{aligned} \sum_{j=1}^N a_j^2 E_\theta(2P_j - 1)^2 = \theta^2 \sum_{j=1}^N a_j^2 E_0\psi_1^2(Z_j) + M_3 , \\ |M_3| \leqq BN\theta^3 ||a^{2r}||^{1/r}[||\psi_3^s|| + ||\psi_2^{3s/2}|| + ||\psi_1^{3s}||]^{1/s} ; \end{aligned} \tag{A1.23}$$

$$\begin{aligned} \sigma_\theta^2\left(\sum_{j=1}^N a_j P_j\right) = \frac{\theta^2}{4} \sigma_0^2\left(\sum_{j=1}^N a_j \psi_1(Z_j)\right) + M_4 , \\ |M_4| \leqq BN^2\theta^{24/8} ||a^2||[||\psi_3^{4/3}|| + ||\psi_2^3|| + ||\psi_1^6||] + BN\theta^{19/8} ||a^3||^{\frac{1}{3}} \\ \times [||\psi^{\frac{4}{3}}|| + ||\psi_2^3|| + ||\psi_1^6||]^{\frac{2}{3}}[E_0|\sum a_j(\psi_1(Z_j) - E_0\psi_1(Z_j))|^3]^{\frac{1}{3}} . \end{aligned} \tag{A1.24}$$

Moreover, for $m \geqq 1$ and $\rho > 0$ there exist B' and B'' depending only on m and on m and ρ respectively, and such that

$$\sum_{j=1}^N E_\theta|2P_j - 1|^m \leqq B'N\theta^m ||\psi_1^m|| ; \tag{A1.25}$$

$$\begin{aligned} \left[\sum_{j=1}^N \{E_\theta|P_j - \pi_j|^m\}^\rho\right]^{1/\rho} \\ \leqq \theta^m\left[\sum \{E_0|\psi_1(Z_j) - E_0\psi_1(Z_j)|^m\}^\rho\right]^{1/\rho} \\ + B''N^{1/\rho}\theta^{2m}[||\psi_2^{m(\rho \vee 1)}|| + ||\psi_1^{2m(\rho \vee 1)}|| + 1]^{1/\rho} , \end{aligned} \tag{A1.26}$$

where $\rho \vee 1$ denotes the larger of ρ and 1.

PROOF. In (A1.14) we take $E = E_0$, $q(Z, \theta) = \sum a_j(2\tilde{p}(Z_j, \theta) - 1)$, $k = 3$,

$m = 1$, and find

$$|M_1| \leqq \frac{\theta^4}{4!} \sup \{E_0|2 \sum a_j \tilde{p}_{0,4}(Z_j, \nu\theta)| : 0 \leqq \nu \leqq 1\}$$
$$\leqq \frac{N\theta^4}{12} \,||a^r||^{1/r} \sup \left\{\left[\frac{1}{N} \sum E_0|\tilde{p}_{0,4}(Z_j, \nu\theta)|^s\right]^{1/s} : 0 \leqq \nu \leqq 1\right\},$$

by Hölder's and Ljapunov's inequalities. Since $\sum |\tilde{p}_{0,4}(Z_j, \nu\theta)|^s$ is symmetric in $Z_1, \cdots, Z_N$, we have

$$\frac{1}{N} \sum E_0|\tilde{p}_{0,4}(Z_j, \nu\theta)|^s = E_0|\tilde{p}_{0,4}(|X_1|, \nu\theta)|^s .$$

Now we apply (A1.13) and use the fact that the distribution of ${}_i\tilde{\psi}_j(|X_1|, \nu\theta)$ under $F(x)$ is the same as that of ${}_i\psi_j(|X_1|, \nu\theta)$ under $F(x - \nu\theta)$ to obtain

$$E_0|\tilde{p}_{0,4}(|X_1|, \nu\theta)|^s \leqq b_4{}^s E_{\nu\theta}[\textstyle\sum_{i=1}^2 \{|{}_i\psi_4(|X_1|, \nu\theta)| + |{}_i\psi_3(|X_1|, \nu\theta)|^{\frac{4}{3}}$$
$$+ {}_i\psi_2{}^2(|X_1|, \nu\theta) + {}_i\psi_1{}^4(|X_1|, \nu\theta)\}]^s .$$

Because $s \geqq 1$ and $0 \leqq 2\nu\theta \leqq \varepsilon$ for $0 \leqq \nu \leqq 1$, this implies that

$$E_0|\tilde{p}_{0,4}(|X_1|, \nu\theta)|^s \leqq 8^{s-1} b_4{}^s[||\psi_4{}^s|| + ||\psi_3{}^{4s/3}|| + ||\psi_2{}^{2s}|| + ||\psi_1{}^{4s}||] ,$$

which proves (A1.21).

The proof of (A1.22), (A1.23) and (A1.25) is similar. In each case we can apply (A1.14), taking $q(Z, \theta) = \sum a_j{}^3(2\tilde{p}(Z_j, \theta) - 1)$, $k = 2$, $m = 1$ to prove (A1.22), and $q(Z, \theta) = \sum a_j{}^2(2\tilde{p}(Z_j, \theta) - 1)^2$, $k = 2$, $m = 1$ to prove (A1.23). In (A1.25) the symmetry in $Z_1, \cdots, Z_N$ is already present from the start, so here we use (A1.14) with $q(|X_1|, \theta) = 2\tilde{p}(|X_1|, \theta) - 1$, $k = 0$ and the value of m as in (A1.25).

A rather delicate argument is needed to deal with (A1.24). Because $\tilde{p}_{0,2}(x, 0) = 0$,

$$\left(\tilde{p}(x, t) - \tfrac{1}{2} + \frac{t}{2} \psi_1(x)\right)^2$$
$$= \left|\frac{t^2}{2} \textstyle\int_0^1 2(1 - \nu)\tilde{p}_{0,2}(x, \nu t)\, d\nu\right|^{\frac{6}{5}} \left|\frac{t^3}{6} \textstyle\int_0^1 3(1 - \nu)^2\tilde{p}_{0,3}(x, \nu t)\, d\nu\right|^{\frac{4}{5}}$$
$$\leqq |t|^{\frac{24}{5}}\{|\tfrac{1}{2} \textstyle\int_0^1 2(1 - \nu)\tilde{p}_{0,2}(x, \nu t)\, d\nu|^3 + |\tfrac{1}{6} \textstyle\int_0^1 3(1 - \nu)^2\tilde{p}_{0,3}(x, \nu t)\, d\nu|^{\frac{4}{3}}\}$$
$$\leqq |t|^{\frac{24}{5}} \textstyle\int_0^1 \{|\tilde{p}_{0,2}(x, \nu t)|^3 + |\tilde{p}_{0,3}(x, \nu t)|^{\frac{4}{3}}\}\, d\nu ,$$

and similarly,

$$\left|\tilde{p}(x, t) - \tfrac{1}{2} + \frac{t}{2} \psi_1(x)\right|^{\frac{3}{2}} \leqq |t|^{\frac{21}{5}} \textstyle\int_0^1 \{|\tilde{p}_{0,2}(x, \nu t)|^3 + |\tilde{p}_{0,3}(x, \nu t)|^{\frac{4}{3}}\}\, d\nu .$$

By now familiar manipulations yield

$$\left|\sigma_\theta{}^2(\textstyle\sum a_j P_j) - \frac{\theta^2}{4} \sigma_0{}^2(\textstyle\sum a_j \psi_1(Z_j))\right|$$
$$\leqq \sigma_0{}^2\left(\textstyle\sum a_j \left\{\tilde{p}(Z_j, \theta) + \frac{\theta}{2} \psi_1(Z_j)\right\}\right)$$
$$+ \theta \left|\mathrm{Cov}_0\left(\textstyle\sum a_j \left\{\tilde{p}(Z_j, \theta) + \frac{\theta}{2} \psi_1(Z_j)\right\} , \sum a_j \psi_1(Z_j)\right)\right|$$

$$\leqq N^2\|a^2\|E_0\left\{\tilde p(|X_1|,\theta)-\tfrac12+\frac{\theta}{2}\psi_1(|X_1|)\right\}^2+N\theta\|a^3\|^{\frac13}\left[E_0\left|\tilde p(|X_1|,\theta)-\tfrac12\right.\right.$$
$$\left.\left.+\frac{\theta}{2}\psi_1(|X_1|)\right|^{\frac32}\right]^{\frac23}[E_0|\textstyle\sum a_j(\psi_1(Z_j)-E_0\psi_1(Z_j))|^3]^{\frac13}$$
$$\leqq BN^2\theta^{\frac{21}{8}}\|a^2\|[\|\psi_3^{\frac43}\|+\|\psi_2^{3}\|+\|\psi_1^{6}\|]+BN\theta^{\frac{19}{8}}\|a^3\|^{\frac13}$$
$$\times[\|\psi_3^{\frac43}\|+\|\psi_2^{3}\|+\|\psi_1^{6}\|]^{\frac23}[E_0|\textstyle\sum a_j(\psi_1(Z_j)-E_0\psi_1(Z_j))|^3]^{\frac13}\,.$$

It remains to consider (A1.26). Since

$$\tilde p(Z_j,\theta)-E_0\tilde p(Z_j,\theta)=\theta[\tilde p_{0,1}(Z_j,0)-E_0\tilde p_{0,1}(Z_j,0)]$$
$$+\frac{\theta^2}{2}\int_0^1[|\tilde p_{0,2}(Z_j,\nu\theta)|+E_0|\tilde p_{0,2}(Z_j,\nu\theta)|]2(1-\nu)\,d\nu\,,$$

and $m\geqq 1$, we have

$$E_\theta|P_j-\pi_j|^m\leqq 2^{m-1}\theta^m E_0|\tilde p_{0,1}(Z_j,0)-E_0\tilde p_{0,1}(Z_j,0)|^m$$
$$+\frac{\theta^{2m}}{2}E_0\int_0^1\{|\tilde p_{0,2}(Z_j,\nu\theta)|+E_0|\tilde p_{0,2}(Z_j,\nu\theta)|^m\}2(1-\nu)\,d\nu$$
$$\leqq\frac{\theta^m}{2}E_0|\psi_1(Z_j)-E_0\psi_1(Z_j)|^m$$
$$+2^{m-1}\theta^{2m}\int_0^1 E_0|\tilde p_{0,2}(Z_j,\nu\theta)|^m2(1-\nu)\,d\nu\,.$$

Hence

$$\sum\{E_\theta|P_j-\pi_j|^m\}^\rho\leqq\theta^{m\rho}\sum\{E_0|\psi_1(Z_j)-E_0\psi_1(Z_j)|^m\}^\rho$$
$$+2^{m\rho}N\theta^{2m\rho}[1+\sup\{E_0|\tilde p_{0,2}(|X_1|,\nu\theta)|^{m(\rho\vee1)}:0\leqq\nu\leqq1\}]\,.$$

Proceeding as before we prove (A1.26) and the theorem. □

COROLLARY A1.1. *Suppose that positive numbers c, C and ε exist such that* (2.35), (3.2) *and* (3.3) *are satisfied. Let $\tilde K$, K_θ and η be defined by* (2.40), (3.4) *and* (3.5). *Then there exists $A>0$ depending on N, a, F and θ only through c, C and ε, and such that*

$$\text{(A1.27)}\qquad \sup_x\left|\tilde K\left(x-\frac{\sum a_j(2\pi_j-1)}{(\sum a_j^2)^{\frac12}}\right)-K_\theta(x-\eta)\right|$$
$$\leqq A\{N^{-\frac54}+\theta^{\frac{19}{8}}[E_0|\textstyle\sum a_j(\psi_1(Z_j)-E_0\psi_1(Z_j))|^3]^{\frac13}$$
$$+N^{-\frac12}\theta^5\sigma_0^2(\textstyle\sum a_j\,\psi_1(Z_j))\}\,,$$

$$\text{(A1.28)}\qquad |\textstyle\sum a_j^m E_0\psi_1(Z_j)|\leqq AN\qquad \textit{for}\quad m=1,3\,,$$

$$\text{(A1.29)}\qquad |\textstyle\sum a_j^2E_0\psi_1^2(Z_j)|\leqq AN\,,$$

$$\text{(A1.30)}\qquad |\textstyle\sum a_jE_0[\psi_3(Z_j)-6\psi_1(Z_j)\psi_2(Z_j)+3\psi_1^3(Z_j)]|\leqq AN\,,$$

$$\text{(A1.31)}\qquad \textstyle\sum E_\theta|2P_j-1|^m\leqq AN^{1-m/2}\qquad \textit{for}\quad 1\leqq m\leqq 6\,,$$

$$\text{(A1.32)}\qquad [\textstyle\sum\{E_\theta|P_j-\pi_j|^3\}^{\frac43}]^{\frac34}\leqq\theta^3[\sum\{E_0|\psi_1(Z_j)-E_0\psi_1(Z_j)|^3\}^{\frac43}]^{\frac34}+AN^{-\frac34}\,.$$

PROOF. Since the corollary is trivially true for $N\leqq(2C/\varepsilon)^2$, we may assume that $2\theta\leqq 2CN^{-\frac12}\leqq\varepsilon$ and use the results in Theorem A1.1. We note that (2.35) implies that $\|a^r\|\leqq[C^r\max(1,N^{r-4})]^{\frac14}$. In the notation of this appendix (3.2) asserts that $\|\psi_i^{m_i}\|\leqq C$ for $m_1=6$, $m_2=3$, $m_3=\frac43$ and $m_4=1$. All order symbols in this proof are uniform for fixed c, C and ε.

(A1.28)—(A1.30) follow from (2.35) and (3.2) by Hölder's and Ljapunov's inequalities, e.g.

$$|\sum a_j^3 E_0 \psi_1(Z_j)| = O(N \|a\|_4^3 \|\psi_1^4\|^{\frac14}) = O(N)\,.$$

(A1.31) and (A1.32) are immediate consequences of (A1.25) and (A1.26).

Taking $r = 4$, $s = \frac{4}{3}$ in (A1.22)—(A1.24) we find

$$\text{(A1.33)} \qquad M_2 = O(1)\,, \qquad M_3 = O(N^{-\frac14})\,,$$
$$M_4 = O(N^{-\frac25} + N\theta^{\frac{19}{8}}[E_0|\sum a_j(\psi_1(Z_j) - E_0\psi_1(Z_j))|^3]^{\frac13})\,.$$

Hence, uniformly in x,

$$\tilde{K}(x) = \Phi(x) + \phi(x)\left\{\frac{\sum a_j^4}{12(\sum a_j^2)^2}(x^3 - 3x) - \theta\frac{\sum a_j^3 E_0\psi_1(Z_j)}{3(\sum a_j^2)^{\frac32}}(x^2 - 1)\right.$$
$$\text{(A1.34)} \qquad \left. + \frac{\theta^2}{2\sum a_j^2}[\sum a_j^2 E_0\psi_1^2(Z_j) - \sigma_0^2(\sum a_j\psi_1(Z_j))]x\right\}$$
$$+ O(N^{-\frac54} + \theta^{\frac{19}{8}}[E_0|\sum a_j(\psi_1(Z_j) - E_0\psi_1(Z_j))|^3]^{\frac13})\,.$$

Taking $r = \infty$, $s = 1$ in (A1.21) we have

$$\text{(A1.35)} \qquad \frac{\sum a_j(2\pi_j - 1)}{(\sum a_j^2)^{\frac12}} = \eta - \frac{\theta^3}{6(\sum a_j^2)^{\frac12}}\sum a_j E_0[\psi_3(Z_j)$$
$$- 6\psi_1(Z_j)\psi_2(Z_j) + 3\psi_1^3(Z_j)] + O(N^{\frac34}\theta^4)\,,$$

where the second term on the right is $O(N^{\frac12}\theta^3)$ by (A1.30). Now we substitute $x - (\sum a_j^2)^{-\frac12}\sum a_j(2\pi_j - 1)$ for x in (A1.34) and expand the right-hand side around $x - \eta$. It follows from (A1.35), (A1.28) for $m = 3$ and (A1.29) that in this way we obtain (A1.27).

2. Asymptotic behavior of moments of functions of order statistics. Our aim in this appendix is twofold. In the first place we provide a proof of Theorem 3.2 where the order of the remainder in expansion (3.8) is evaluated. Secondly, we obtain asymptotic expressions for the leading terms in the expansion for the case where exact or approximate scores are used, thus in effect proving Theorems 4.1 and 4.2.

Let $U_{1:N} < U_{2:N} < \cdots < U_{N:N}$ be order statistics of a sample of size N from the uniform distribution on (0, 1).

LEMMA A2.1. *If $\lambda = j/(N+1)$ then for all $N = 1, 2, \cdots, j = 1, \cdots, N$ and $t \geqq 0$,*

$$P\left(|U_{j:N} - \lambda|\left(\frac{N}{\lambda(1-\lambda)}\right)^{\frac12} \geqq t\right) \leqq 2\exp\left\{-\frac{3t^2}{6t+8}\right\}.$$

PROOF. The probability on the left is equal to

$$\text{(A2.1)} \qquad B\left(j, N, \lambda - t\left(\frac{\lambda(1-\lambda)}{N}\right)^{\frac12}\right)$$
$$+ B\left(N - j + 1, N, 1 - \lambda - t\left(\frac{\lambda(1-\lambda)}{N}\right)^{\frac12}\right)$$

where

$$B(j, N, p) = \sum_{k=j}^{N}\binom{N}{k}p^k(1-p)^{N-k}\,.$$

For $j > Np$ Bernstein's inequality (cf. Hoeffding (1963) page 17) yields

$$B(j, N, p) \leqq \exp\left\{-\frac{j - Np}{1 - p}\, h\left(\frac{j - Np}{Np}\right)\right\}$$

with $h(s) = 3s(2s + 6)^{-1}$. Application of this result gives after some algebra

$$B\left(j, N, \lambda - t\left(\frac{\lambda(1-\lambda)}{N}\right)^{\frac{1}{2}}\right)$$
$$\leqq \exp\left\{-\frac{3}{2}\,\frac{[t + (\lambda/N(1-\lambda))^{\frac{1}{2}}]^2}{(3 + N^{-1}) + t(N\lambda(1-\lambda))^{-\frac{1}{2}}[\lambda(5 + N^{-1}) - 2] - 2N^{-1}t^2}\right\}.$$

Noting that $\lambda \leqq N(N+1)^{-1}$ and $(N\lambda(1-\lambda))^{-\frac{1}{2}} \leqq 1 + N^{-1}$, we see that $\exp\{-3t^2(6t+8)^{-1}\}$ is an upper bound for the first term in (A2.1). By interchanging j and $(N - j + 1)$ we find that the same is true for the second term in (A2.1) which proves the lemma. □

LEMMA A2.2. *If $\lambda = j/(N+1)$, k is a positive real number, ν_k is the kth absolute moment of the standard normal distribution and $I_{(a,b)}$ is the indicator of (a, b), then uniformly for $j = 1, \cdots, N$ and $\eta \geqq \frac{1}{2}\lambda(1-\lambda)$ we have for $N \to \infty$,*

$$\left(\frac{N}{\lambda(1-\lambda)}\right)^{\frac{1}{2}k} E(\lambda - U_{j:N})^k I_{(\lambda-\eta,\lambda)}(U_{j:N}) = \tfrac{1}{2}\nu_k + O((N\lambda(1-\lambda))^{-\frac{1}{2}}),$$

$$\left(\frac{N}{\lambda(1-\lambda)}\right)^{\frac{1}{2}k} E(U_{j:N} - \lambda)^k I_{(\lambda,\lambda+\eta)}(U_{j:N}) = \tfrac{1}{2}\nu_k + O((N\lambda(1-\lambda))^{-\frac{1}{2}}).$$

PROOF. Let f be the density of $Z = (N/\lambda(1-\lambda))^{\frac{1}{2}}(U_{j:N} - \lambda)$. Application of Stirling's formula in the form $\log n! = (n + \frac{1}{2})\log(n+1) - (n+1) + \frac{1}{2}\log 2\pi + O(n^{-1})$ followed by expansion of logarithms yields

$$\log f(z) = -\tfrac{1}{2}\log 2\pi + \frac{2\lambda - 1}{(N\lambda(1-\lambda))^{\frac{1}{2}}}\, z - \frac{1}{2}\left[1 - \frac{\lambda^3 + (1-\lambda)^3}{N\lambda(1-\lambda)}\right] z^2$$
$$+ O\left(\frac{|z|^3}{(N\lambda(1-\lambda))^{\frac{1}{2}}} + \frac{1}{N\lambda(1-\lambda)}\right)$$

for $z^2 < N\min(\lambda/(1-\lambda), (1-\lambda)/\lambda)$. Hence, for $|z| \leqq (N\lambda(1-\lambda))^{\frac{1}{6}} < [N\min(\lambda/(1-\lambda), (1-\lambda)/\lambda)]^{\frac{1}{2}}$,

$$f(z) = \frac{1}{(2\pi)^{\frac{1}{2}}} e^{-\frac{1}{2}z^2}\left[1 + O\left(\frac{|z| + |z|^3}{(N\lambda(1-\lambda))^{\frac{1}{2}}} + \frac{1}{N\lambda(1-\lambda)}\right)\right] \tag{A2.2}$$

uniformly in j. Since $\eta(N/\lambda(1-\lambda))^{\frac{1}{2}} \geqq \frac{1}{2}(N\lambda(1-\lambda))^{\frac{1}{6}}$, (A2.2) and Lemma A2.1 imply that

$$EZ^k I_{(\lambda,\lambda+\eta)}(U_{j:N}) = \frac{1}{(2\pi)^{\frac{1}{2}}} \int_0^{\frac{1}{2}(N\lambda(1-\lambda))^{\frac{1}{6}}} z^k e^{-\frac{1}{2}z^2}\left[1 + O\left(\frac{1 + |z| + |z|^3}{(N\lambda(1-\lambda))^{\frac{1}{2}}}\right)\right] dz$$
$$+ O\left(\int_{\frac{1}{2}(N\lambda(1-\lambda))^{\frac{1}{6}}}^{\infty} z^k e^{-\frac{1}{2}z} dz\right) = \tfrac{1}{2}\nu_k + O((N\lambda(1-\lambda))^{-\frac{1}{2}}),$$

which proves the second part of the lemma. The first part now follows by noting that $U_{j:N}$ and $1 - U_{N-j+1:N}$ have the same distribution. □

REMARK. One easily verifies that Lemma A2.2 continues to hold when η is taken as small as $[c(\lambda(1-\lambda)/N)|\log N\lambda(1-\lambda)|]^{\frac{1}{2}}$ for any $c > 1$. It should also be noted that when j or $(N-j+1)$ remains bounded as $N \to \infty$, Lemma A2.2 merely states that $E|U_{j:N} - \lambda|^k = O(N^{-k})$.

Condition R_r. For real $r > 0$, a function h on $(0, 1)$ is said to satisfy condition R_r if h is twice continuously differentiable on $(0, 1)$ and

$$\limsup_{t\to 0,1} t(1-t)\left|\frac{h''(t)}{h'(t)}\right| < 1 + \frac{1}{r}\,.$$

LEMMA A2.3. *Let $r_1, \cdots, r_m, k_1, \cdots, k_m$ be positive real numbers, $j = 1, \cdots, N$, $\lambda = j/(N+1)$ and ν_k the kth absolute moment of the standard normal distribution. Suppose that $h_1, \cdots, h_m$ satisfy conditions $R_{r_1}, \cdots, R_{r_m}$ respectively and that $\sum k_i/r_i \leqq 1$. Define*

$$M = \left(\frac{\lambda(1-\lambda)}{N}\right)^{\frac{1}{2}\sum k_i} \left\{\left(\frac{\lambda(1-\lambda)}{N}\right)^{\frac{1}{2}} + (N\lambda(1-\lambda))^{-\frac{1}{2}} \prod_{i=1}^m |h_i'(\lambda)|^{k_i}\right\}.$$

Then, uniformly in j, we have for $N \to \infty$

$$E \prod_{i=1}^m |h_i(U_{j:N}) - h_i(\lambda)|^{k_i} = \left(\frac{\lambda(1-\lambda)}{N}\right)^{\frac{1}{2}\sum k_i} \nu_{\sum k_i} \prod_{i=1}^m |h_i'(\lambda)|^{k_i} + O(M)$$

and for integer $k_1, \cdots, k_m$

$$\begin{aligned} E \prod_{i=1}^m &(h_i(U_{j:N}) - h_i(\lambda))^{k_i} \\ &= O(M) \qquad \textit{if} \quad \textstyle\sum k_i \quad \textit{is odd}, \\ &= \left(\frac{\lambda(1-\lambda)}{N}\right)^{\frac{1}{2}\sum k_i} \nu_{\sum k_i} \prod_{i=1}^m (h_i'(\lambda))^{k_i} + O(M) \qquad \textit{if} \quad \textstyle\sum k_i \quad \textit{is even}. \end{aligned}$$

PROOF. For reasons of symmetry it is sufficient to consider only $j \leqq (N+1)/2$, i.e. $\lambda \leqq \frac{1}{2}$. Since h_i satisfies condition R_{r_i}, there exists $0 < \varepsilon < \frac{1}{6}$, $\tau > 1$ and $C > 0$ such that for $i = 1, \cdots, m$

$$\text{(A2.3)} \qquad \left|\frac{h_i''(t)}{h_i'(t)}\right| \leqq \left(1 + \frac{1}{r_i\tau}\right) t^{-1} \qquad \text{for} \quad 0 < t \leqq 3\varepsilon\,,$$

$$\text{(A2.4)} \qquad |h_i''(t)| \leqq C \qquad \text{for} \quad \varepsilon \leqq t \leqq 1 - \varepsilon\,,$$

$$\text{(A2.5)} \qquad \left|\frac{h_i''(t)}{h_i'(t)}\right| \leqq \left(1 + \frac{1}{r_i\tau}\right)(1-t)^{-1} \qquad \text{for} \quad 1 - 3\varepsilon \leqq t < 1\,.$$

Suppose first that $\lambda \leqq 2\varepsilon$. Integration of (A2.3) shows that for $0 < t \leqq \lambda$ and $i = 1, \cdots, m$,

$$\left(\frac{t}{\lambda}\right)^{1+1/r_i\tau} \leqq \frac{h_i'(t)}{h_i'(\lambda)} \leqq \left(\frac{\lambda}{t}\right)^{1+1/r_i\tau},$$

$$\frac{r_i\tau}{2r_i\tau + 1}\,\lambda\left[1 - \left(\frac{t}{\lambda}\right)^{2+1/r_i\tau}\right] \leqq \frac{h_i(\lambda) - h_i(t)}{h_i'(\lambda)} \leqq r_i\tau\lambda\left[\left(\frac{\lambda}{t}\right)^{1/r_i\tau} - 1\right].$$

It follows that

$$\text{(A2.6)} \qquad \frac{h_i(\lambda) - h_i(t)}{h_i'(\lambda)} = (\lambda - t) + O\left(\frac{(\lambda - t)^2}{\lambda}\right) \qquad \text{for} \quad \tfrac{1}{2}\lambda \leqq t \leqq \lambda\,,$$

$$\left|\frac{h_i(\lambda) - h_i(t)}{h_i'(\lambda)}\right| \leqq r_i\tau\lambda\left(\frac{\lambda}{t}\right)^{1/r_i\tau} \qquad \text{for} \quad 0 < t \leqq \tfrac{1}{2}\lambda\,.$$

Application of Lemma A2.2 with $\eta = \frac{1}{2}\lambda$ yields

$$
\begin{aligned}
&E \prod_{i=1}^{m} \left(\frac{h_i(\lambda) - h_i(U_{j:N})}{h_i'(\lambda)}\right)^{k_i} I_{(0,\lambda)}(U_{j:N}) \\
\text{(A2.7)} \qquad &= \frac{1}{2}\left(\frac{\lambda(1-\lambda)}{N}\right)^{\frac{1}{2}\Sigma k_i} \nu_{\Sigma k_i}[1 + O((N\lambda(1-\lambda))^{-\frac{1}{2}})] \\
&\quad + O\left(\lambda^{\Sigma k_i} E\left(\frac{\lambda}{U_{j:N}}\right)^{1/\tau} I_{(0,\frac{1}{2}\lambda)}(U_{j:N})\right),
\end{aligned}
$$

where we have made use of $\sum k_i/r_i \leqq 1$. For $2 \leqq j \leqq \frac{1}{2}(N+1)$,

$$
\begin{aligned}
\lambda^{\Sigma k_i} E\left(\frac{\lambda}{U_{j:N}}\right)^{1/\tau} I_{(0,\frac{1}{2}\lambda)}(U_{j:N}) &= \lambda^{\Sigma k_i + 1/\tau} \frac{N}{j-1} E U_{j-1:N-1}^{1-1/\tau} I_{(0,\frac{1}{2}\lambda)}(U_{j-1:N-1}) \\
\text{(A2.8)} \qquad &\leqq 2\lambda^{\Sigma k_i} P(U_{j-1:N-1} < \tfrac{1}{2}\lambda) \\
&= O\left(\left(\frac{\lambda(1-\lambda)}{N}\right)^{\frac{1}{2}\Sigma k_i} (N\lambda(1-\lambda))^{-\frac{1}{2}}\right)
\end{aligned}
$$

by Lemma A2.1. For $j = 1$ we have

$$
\begin{aligned}
&\lambda^{\Sigma k_i} E\left(\frac{\lambda}{U_{j:N}}\right)^{1/\tau} I_{(0,\frac{1}{2}\lambda)}(U_{j:N}) \\
\text{(A2.9)} \qquad &= (N+1)^{-\Sigma k_i - 1/\tau} N \int_0^{1/2(N+1)} u^{-1/\tau}(1-u)^{N-1}\, du \\
&= O(N^{-\Sigma k_i}) = O\left(\left(\frac{\lambda(1-\lambda)}{N}\right)^{\frac{1}{2}\Sigma k_i} (N\lambda(1-\lambda))^{-\frac{1}{2}}\right).
\end{aligned}
$$

Together, (A2.8) and (A2.9) ensure that the second remainder term in (A2.7) may be omitted.

A similar analysis based on (A2.3)—(A2.5) shows that for $\lambda \leqq 2\varepsilon$ but $t \geqq \lambda$, (A2.6) holds for $\lambda \leqq t \leqq 3\lambda/2$ and

$$
\begin{aligned}
\left|\frac{h_i(t) - h_i(\lambda)}{h_i'(\lambda)}\right| &\leqq r_i \tau \lambda \left(\frac{t}{\lambda}\right)^{2+1/r_i\tau} \qquad \text{for} \quad \frac{3\lambda}{2} \leqq t \leqq 3\varepsilon, \\
&= O(\lambda^{-1-1/r_i\tau}(1-t)^{-1/r_i\tau}) \qquad \text{for} \quad 3\varepsilon \leqq t < 1.
\end{aligned}
$$

Hence by Lemmas A2.2 and A2.1 and a change from $U_{j:N}$ to $U_{j:N-1}$ as in (A2.8),

$$
\begin{aligned}
&E \prod_{i=1}^{m} \left(\frac{h_i(U_{j:N}) - h_i(\lambda)}{h_i'(\lambda)}\right)^{k_i} I_{(\lambda,1)}(U_{j:N}) \\
&= \frac{1}{2}\left(\frac{\lambda(1-\lambda)}{N}\right)^{\frac{1}{2}\Sigma k_i} \nu_{\Sigma k_i}[1 + O((N\lambda(1-\lambda))^{-\frac{1}{2}})] \\
\text{(A2.10)} \qquad &\quad + O(\lambda^{\Sigma k_i} \exp\{-\tfrac{1}{4}(N\lambda)^{\frac{1}{2}}\} \\
&\quad + \lambda^{-\Sigma k_i - 1/\tau} E(1 - U_{j:N-1})^{1-1/\tau} I_{(3\varepsilon,1)}(U_{j:N-1})) \\
&= \frac{1}{2}\left(\frac{\lambda(1-\lambda)}{N}\right)^{\frac{1}{2}\Sigma k_i} \nu_{\Sigma k_i}[1 + O((N\lambda(1-\lambda))^{-\frac{1}{2}})].
\end{aligned}
$$

Combining (A2.7)—(A2.10) and noting that (A2.7) and (A2.10) remain valid when absolute values are taken inside the expectation signs, we see that the lemma is proved for $\lambda \leqq 2\varepsilon$.

If $2\varepsilon < \lambda \leqq \frac{1}{2}$, (A2.3)—(A2.5) imply that

$$h_i(t) - h_i(\lambda) = h_i'(\lambda)(t-\lambda) + O((t-\lambda)^2) \qquad \text{for} \quad \varepsilon \leqq t \leqq 1-\varepsilon\,,$$
$$|h_i(t) - h_i(\lambda)| = O((t(1-t))^{-1/r_i\tau}) \qquad \text{for} \quad t < \varepsilon \quad \text{or} \quad t > 1-\varepsilon\,,$$

and the proof of the lemma for $2\varepsilon < \lambda \leqq \frac{1}{2}$ follows by noting that $h_i'(\lambda)$ is bounded and arguing as e.g. in (A2.10). □

REMARK. Although the remainder M in Lemma A2.3 consists of two terms, only one of these plays a role for any particular value of λ. For $2\varepsilon < \lambda < 1 - 2\varepsilon$, $h_i'(\lambda)$ and $(\lambda(1-\lambda))^{-1}$ are bounded and we need only retain the first term of M. It follows from (A2.7)—(A2.10) that for $\lambda \leqq 2\varepsilon$ or $\lambda \geqq 1 - 2\varepsilon$ only the second term of M is needed.

LEMMA A2.4. *Lemma* A2.3 *continues to hold for central moments,* i.e. *if $h_i(\lambda)$ is replaced by $Eh_i(U_{j:N})$ for $i = 1, \cdots, m$, provided only that $r_i \geqq 1$ for $i = 1, \cdots, m$.*

PROOF. As $r_i \geqq 1$, Lemma A2.3 contains as a special case

$$|Eh_i(U_{j:N}) - h_i(\lambda)| = O\left(\frac{\lambda(1-\lambda) + |h_i'(\lambda)|}{N}\right). \tag{A2.11}$$

The lemma is proved by expanding the central moments in terms of moments centered at the $h_i(\lambda)$ and applying (A2.11), Lemma A2.3 and the remark following it. □

We also note the following extension of a result of Hoeffding (1953).

LEMMA A2.5. *Let $h_1, \cdots, h_m$ be continuous functions on* (0, 1), *q a continuous function on R^m and Q a convex function on R^m such that $|q| \leqq Q$. Suppose that $\int_0^1 |h_i(t)|\,dt < \infty$ for $i = 1, \cdots, m$ and that $\int_0^1 Q(h_1(t), \cdots, h_m(t))\,dt < \infty$. Then*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^N q(Eh_1(U_{j:N}), \cdots, Eh_m(U_{j:N})) = \int_0^1 q(h_1(t), \cdots, h_m(t))\,dt\,.$$

PROOF. Because h_i is continuous and summable, Lemma 2.2 of Bickel (1967) implies that for any $\varepsilon > 0$, $Eh_i(U_{j_N:N}) - h_i(j_N(N+1)^{-1}) \to 0$ uniformly for $\varepsilon \leqq j_N(N+1)^{-1} \leqq 1 - \varepsilon$ as $N \to \infty$. Since q is continuous and $q(h_1, \cdots, h_m)$ is summable, the lemma is proved if we show that

$$\lim_{\varepsilon\downarrow 0} \limsup_N \frac{1}{N} \left(\sum_{j=1}^{[\varepsilon(N+1)]} + \sum_{j=[(1-\varepsilon)(N+1)]}^N\right)|q(Eh_1(U_{j:N}), \cdots, Eh_m(U_{j:N}))| = 0\,.$$

It is obviously sufficient to prove this for Q instead of q, but as Q has the same properties as q and is moreover nonnegative, this is equivalent to showing that

$$\limsup_N \frac{1}{N} \sum_{j=1}^N Q(Eh_1(U_{j:N}), \cdots, Eh_m(U_{j:N})) \leqq \int_0^1 Q(h_1(t), \cdots, h_m(t))\,dt\,.$$

As Q is convex this follows from Jensen's inequality. □

LEMMA A2.6. *Let $k_1, \cdots, k_m$ be positive integers and $r_1, \cdots, r_m$ positive real numbers such that $\sum k_i/r_i \leqq 1$. Suppose that $h_1, \cdots, h_m$ are continuous functions* *on* $(0, 1)$ *for which* $\int_0^1 |h_i(t)|^{r_i}\, dt < \infty$ *for* $i = 1, \cdots, m$. *Then*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^N \prod_{i=1}^m (Eh_i(U_{j:N}))^{k_i} = \int_0^1 \prod_{i=1}^m (h_i(t))^{k_i}\, dt\,.$$

If, in addition, h_1 *is monotone in neighborhoods of* 0 *and* 1, *then also*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^N \left(h_1\left(\frac{j}{N+1}\right)\right)^{k_1} \prod_{i=2}^m (Eh_i(U_{j:N}))^{k_i} = \int_0^1 \prod_{i=1}^m (h_i(t))^{k_i}\, dt\,.$$

PROOF. The first part of the lemma is a special case of Lemma A2.5, obtained by taking $q(x_1, \cdots, x_m) = \prod x_i^{k_i}$ and $Q(x_1, \cdots, x_m) = 1 + \sum |x_i|^{r_i}$. To establish the second part we follow the proof of Lemma A2.5 for these choices of q and Q but with $Eh_1(U_{j:N})$ replaced by $h_1(j(N+1)^{-1})$, until we arrive at the point where it suffices to show that

$$\limsup_N \frac{1}{N} \sum_{j=1}^N \left[\left|h_1\left(\frac{j}{N+1}\right)\right|^{r_1} + \sum_{i=2}^m |Eh_i(U_{j:N})|^{r_i}\right] \leqq \int_0^1 \sum_{i=1}^m |h_i(t)|^{r_i}\, dt\,.$$

As $|h_1|^{r_1}$ is continuous and summable, its monotonicity near 0 and 1 amply guarantees that $N^{-1} \sum |h_1(j(N+1)^{-1})|^{r_1} \to \int_0^1 |h_1(t)|^{r_1}\, dt$. Application of Jensen's inequality to the remaining terms completes the proof. □

We now state the results needed to prove Theorems 3.2, 4.1 and 4.2 in the form of two corollaries.

COROLLARY A2.1. *Suppose that positive numbers* C *and* δ *exist such that* $|h'(t)| \leqq C(t(1-t))^{-\frac{4}{3}+\delta}$ *for all* $0 < t < 1$. *Then there exists* $A > 0$ *depending on* N *and* h *only through* C *and* δ *and such that*

$$\sum_{j=1}^N \{E|h(U_{j:N}) - Eh(U_{j:N})|^3\}^{\frac{4}{9}} \leqq AN^{\frac{4}{9}}\,.$$

The above condition is fulfilled if h *satisfies condition* R_1 *and* $\int_0^1 h^6(t)\, dt < \infty$.

PROOF. Define $\lambda = j/(N+1)$. For all $0 < t < 1$, $|h(t) - h(\lambda)|$ is maximized by taking $h'(t) \equiv C(t(1-t))^{-\frac{4}{3}+\delta}$ and for this particular choice of h' the function h satisfies condition R_3. Hence, by Lemma A2.3, we have in general

$$E|h(U_{j:N}) - h(\lambda)|^k = O\left(\left(\frac{\lambda(1-\lambda)}{N}\right)^{\frac{1}{2}k} (\lambda(1-\lambda))^{-k(\frac{4}{3}-\delta)}\right)$$

for $0 < k \leqq 3$. It follows that

$$\begin{aligned}\sum_{j=1}^N \{E|h(U_{j:N}) - Eh(U_{j:N})|^3\}^{\frac{4}{9}} &= O(\textstyle\sum_{j=1}^N \{N^{-\frac{3}{2}}(\lambda(1-\lambda))^{-\frac{5}{2}}\}^{\frac{4}{9}}) \\ &= O(N^{\frac{1}{3}} \int_{1/N}^{1-1/N} (t(1-t))^{-\frac{10}{9}}\, dt) = O(N^{\frac{4}{9}})\,.\end{aligned}$$

Condition R_1 ensures that for ε as in (A2.3) and $0 < t < \frac{1}{2}u < \varepsilon$, $|h(t) - h(2\varepsilon)| \geqq \frac{1}{4}u|h'(u)|$ and hence for $u \to 0$,

$$u^7(h'(u))^6 \leqq 2^{13} \int_0^{\frac{1}{2}u} (h(t) - h(2\varepsilon))^6\, dt \to 0\,.$$

In the same way one shows that $|h'(u)| = o((1-u)^{-\frac{7}{6}})$ for $u \to 1$, which completes the proof. □

For $i = 1, 2, 3$, let $\psi_i = f^{(i)}/f$ and $\Psi_i(t) = \psi_i(F^{-1}((1+t)/2))$ as in (3.1) and (3.15). Let J be a function on (0, 1).

COROLLARY A2.2. *Suppose that* (3.2) *holds, that* $0 < \int_0^1 J^4(t)\,dt < \infty$ *and that both* J *and* Ψ_1 *satisfy condition* R_2. *Let either* $a_j = a_{j,N} = EJ(U_{j:N})$ *for* $j = 1, \cdots, N$ *or* $a_j = a_{j,N} = J(j/(N+1))$ *for* $j = 1, \cdots, N$. *Then, as* $N \to \infty$,

$$\frac{1}{N} \sum_{j=1}^N a_j^2 = \int_0^1 J^2(t)\,dt + o(1)\,, \tag{A2.12}$$

$$\frac{1}{N} \sum_{j=1}^N a_j^k E\Psi_1^{4-k}(U_{j:N}) = \int_0^1 J^k(t)\Psi_1^{4-k}(t)\,dt + o(1)\,, \qquad k = 1, \cdots, 4\,, \tag{A2.13}$$

$$\frac{1}{N} \sum_{j=1}^N a_j E\Psi_1(U_{j:N})\Psi_2(U_{j:N}) = \int_0^1 J(t)\Psi_1(t)\Psi_2(t)\,dt + o(1)\,, \tag{A2.14}$$

$$\frac{1}{N} \sum_{j=1}^N a_j E\Psi_3(U_{j:N}) = \int_0^1 J(t)\Psi_3(t)\,dt + o(1)\,, \tag{A2.15}$$

$$\begin{aligned} &\frac{1}{N} \sigma^2\Big(\sum_{j=1}^N a_j \Psi_1(U_{j:N})\Big) \\ &\quad = \int_0^1 \int_0^1 J(s)J(t)\Psi_1'(s)\Psi_1'(t)[s \wedge t - st]\,ds\,dt + o(1)\,. \end{aligned} \tag{A2.16}$$

If $a_j = EJ(U_{j:N})$ *for* $j = 1, \cdots, N$, *then also*

$$\begin{aligned} &N^{-\frac{1}{2}} \frac{\sum_{j=1}^N a_j E\Psi_1(U_{j:N})}{(\sum_{j=1}^N a_j^2)^{\frac{1}{2}}} \\ &\quad = \frac{\int_0^1 J(t)\Psi_1(t)\,dt}{(\int_0^1 J^2(t)\,dt)^{\frac{1}{2}}} - \frac{1}{N} \frac{\sum_{j=1}^N \operatorname{Cov}(J(U_{j:N}), \Psi_1(U_{j:N}))}{(\int_0^1 J^2(t)\,dt)^{\frac{1}{2}}} \\ &\qquad + \frac{1}{2N} \frac{\int_0^1 J(t)\Psi_1(t)\,dt}{(\int_0^1 J^2(t)\,dt)^{\frac{3}{2}}} \sum_{j=1}^N \sigma^2(J(U_{j:N})) + o(N^{-1}) \\ &\quad = \frac{\int_0^1 J(t)\Psi_1(t)\,dt}{(\int_0^1 J^2(t)\,dt)^{\frac{1}{2}}} - \frac{1}{N} \frac{\int_{1/N}^{1-1/N} J'(t)\Psi_1'(t)t(1-t)\,dt}{(\int_0^1 J^2(t)\,dt)^{\frac{1}{2}}} \\ &\qquad + \frac{1}{2N} \frac{\int_0^1 J(t)\Psi_1(t)\,dt}{(\int_0^1 J^2(t)\,dt)^{\frac{3}{2}}} \int_{1/N}^{1-1/N} (J'(t))^2 t(1-t)\,dt + o(N^{-1}) \\ &\qquad + O\Big(N^{-\frac{3}{2}} \int_{1/N}^{1-1/N} |J'(t)|(|J'(t)| + |\Psi_1'(t)|)(t(1-t))^{\frac{1}{2}}\,dt\Big)\,. \end{aligned} \tag{A2.17}$$

If $J = -\Psi_1$ *and either* $a_j = -E\Psi_1(U_{j:N})$ *for* $j = 1, \cdots, N$ *or* $a_j = -\Psi_1(j/(N+1))$ *for* $j = 1, \cdots, N$, *then*

$$\begin{aligned} &N^{-\frac{1}{2}} \frac{\sum_{j=1}^N a_j E\Psi_1(U_{j:N})}{(\sum_{j=1}^N a_j^2)^{\frac{1}{2}}} \\ &\quad = -\Big(\int_0^1 \Psi_1^2(t)\,dt\Big)^{\frac{1}{2}} + \frac{\int_{1/N}^{1-1/N} (\Psi_1'(t))^2 t(1-t)\,dt}{2N(\int_0^1 \Psi_1^2(t)\,dt)^{\frac{1}{2}}} \\ &\qquad + o(N^{-1}) + O\Big(N^{-\frac{3}{2}} \int_{1/N}^{1-1/N} (\Psi_1'(t))^2 (t(1-t))^{\frac{1}{2}}\,dt\Big)\,. \end{aligned} \tag{A2.18}$$

PROOF. The assumptions imply that Ψ_1, Ψ_2, Ψ_3 and J are continuous, that Ψ_1^6, Ψ_2^3, $|\Psi_3|^{\frac{4}{3}}$ and J^4 are summable and that J is monotone near 0 and 1. Hence (A2.12)—(A2.15) follow from Lemma A2.6.

For $a_j = J(j/(N+1))$ a proof of (A2.16) is essentially contained in Stigler (1969). Our condition R_2 for Ψ_1 ensures that Ψ_1' will satisfy Stigler's condition T at 0 and 1. As in the proof of Corollary A2.1, one can argue that near 0 and 1

$$\text{(A2.19)} \qquad \Psi_1'(t) = o((t(1-t))^{-\frac{7}{6}}) , \qquad J'(t) = o((t(1-t))^{-\frac{5}{4}}) .$$

Inspection of Stigler's conditions for (A2.16) shows that in our case the only missing ingredient is that Ψ_1 is not necessarily increasing on (0, 1). However, Ψ_1 is monotone where it matters, that is in a neighborhood of 0 and 1.

To prove that (A2.16) remains valid for $a_j = EJ(U_{j:N})$ we note that by Lemma A2.4 and (A2.19)

$$\begin{aligned}
&\sigma^2\left(\sum_{j=1}^N \left(EJ(U_{j:N}) - J\left(\frac{j}{N+1}\right)\right) \Psi_1(U_{j:N})\right) \\
&\qquad \leqq \left[\sum_{j=1}^N \left|EJ(U_{j:N}) - J\left(\frac{j}{N+1}\right)\right| \sigma(\Psi_1(U_{j:N}))\right]^2 \\
&\qquad = o(N^{-1}[\textstyle\int_{1/N}^{1-1/N} (t(1-t))^{-\frac{23}{12}}\, dt]^2) = o(N^{\frac{5}{6}}) .
\end{aligned}$$

For $a_j = EJ(U_{j:N})$ we have

$$\text{(A2.20)} \qquad \frac{1}{N} \sum_{j=1}^N a_j^2 = \int_0^1 J^2(t)\, dt - \frac{1}{N} \sum_{j=1}^N \sigma^2(J(U_{j:N})) ,$$

$$\begin{aligned}
\text{(A2.21)} \qquad &\frac{1}{N} \sum_{j=1}^N a_j E\Psi_1(U_{j:N}) \\
&\qquad = \int_0^1 J(t)\Psi_1(t)\, dt - \frac{1}{N} \sum_{j=1}^N \operatorname{Cov}(J(U_{j:N}), \Psi_1(U_{j:N})) .
\end{aligned}$$

By Lemma A2.4, condition R_2 for J, and (A2.19)

$$\begin{aligned}
&\frac{1}{N} \sum_{j=1}^N \sigma^2(J(U_{j:N})) \\
&\qquad = \frac{1}{N} \int_{1/N}^{1-1/N} (J'(t))^2 t(1-t)\, dt + O(N^{-2} \int_{1/N}^{1-1/N} (J'(t))^2\, dt + N^{-\frac{3}{2}} \\
\text{(A2.22)} \qquad &\qquad\qquad + N^{-\frac{3}{2}} \int_{1/N}^{1-1/N} (J'(t))^2 (t(1-t))^{\frac{1}{2}}\, dt) \\
&\qquad = \frac{1}{N} \int_{1/N}^{1-1/N} (J'(t))^2 t(1-t)\, dt \\
&\qquad\qquad + O(N^{-\frac{3}{2}} + N^{-\frac{3}{2}} \int_{1/N}^{1-1/N} (J'(t))^2 (t(1-t))^{\frac{1}{2}}\, dt) = o(N^{-\frac{1}{2}}) .
\end{aligned}$$

Similarly

$$\begin{aligned}
&\frac{1}{N} \sum_{j=1}^N \operatorname{Cov}(J(U_{j:N}), \Psi_1(U_{j:N})) \\
\text{(A2.23)} \qquad &\qquad = \frac{1}{N} \int_{1/N}^{1-1/N} J'(t)\Psi_1'(t) t(1-t)\, dt \\
&\qquad\qquad + O(N^{-\frac{3}{2}} + N^{-\frac{3}{2}} \int_{1/N}^{1-1/N} |J'(t)\Psi_1'(t)|(t(1-t))^{\frac{1}{2}}\, dt) = o(N^{-\frac{7}{12}}) .
\end{aligned}$$

Together (A2.20)—(A2.23) are sufficient to prove (A2.17).

If $J = -\Psi_1$ and $a_j = -E\Psi_1(U_{j:N})$, then (A2.17) reduces to (A2.18). To prove that (A2.18) also holds if $a_j = -\Psi_1(j/(N+1))$, it suffices to show that

$$\begin{aligned} &\textstyle\sum_{j=1}^N \Psi_1\left(\frac{j}{N+1}\right) E\Psi_1(U_{j:N}) \\ \text{(A2.24)} \qquad &- \left[\textstyle\sum_{j=1}^N \Psi_1^2\left(\frac{j}{N+1}\right) \sum_{j=1}^N (E\Psi_1(U_{j:N}))^2\right]^{\frac{1}{2}} \\ &= o(1) + O(N^{-\frac{1}{2}} \textstyle\int_{1/N}^{1-1/N} (\Psi_1'(t))^2(t(1-t))^{\frac{1}{2}}\, dt)\,. \end{aligned}$$

It follows from Lemma A2.3 and condition R_2 for Ψ_1 that

$$\begin{aligned} \textstyle\sum_{j=1}^N \left\{E\Psi_1(U_{j:N}) - \Psi_1\left(\frac{j}{N+1}\right)\right\}^2 &= O(N^{-1} + N^{-1} \textstyle\int_{1/N}^{1-1/N} (\Psi_1'(t))^2\, dt) \\ &= O(N^{-1} + N^{-\frac{1}{2}} \textstyle\int_{1/N}^{1-1/N} (\Psi_1'(t))^2(t(1-t))^{\frac{1}{2}}\, dt)\,, \end{aligned}$$

which suffices to establish (A2.24) and complete the proof. □

REFERENCES

ALBERS, W. (1974). Asymptotic expansions and the deficiency concept in statistics. *Mathematical Centre Tracts* **58**, Amsterdam.

BICKEL, P. J. (1967). Some contributions to the theory of order statistics. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1** 575–591. Univ. of California Press.

BICKEL, P. J. (1974). Edgeworth expansions in nonparametric statistics. *Ann. Statist.* **2** 1–20.

BICKEL, P. J. and VAN ZWET, W. R. (1975). Asymptotic expansions for the power of distribution free tests in the two-sample problem. In preparation.

CHUNG, K. L. (1951). The strong law of large numbers. *Proc. Second Berkeley Symp. Math. Statist. Prob.* 341–352. Univ. of California Press.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics.* Princeton Univ. Press.

CRAMÉR, H. (1962). *Random Variables and Probability Distributions*, 2nd ed. Cambridge Univ. Press.

DIEUDONNÉ, J. (1960). *Foundations of Modern Analysis.* Academic Press, New York.

FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

HODGES, J. L. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34** 598–611.

HODGES, J. L. and LEHMANN, E. L. (1970). Deficiency. *Ann. Math. Statist.* **41** 783–801.

HOEFFDING, W. (1953). On the distribution of the expected values of the order statistics. *Ann. Math. Statist.* **24** 93–100.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

OKAMOTO, M. (1958). Some inequalities relating to the partial sum of binomial probabilities. *Ann. Inst. Statist. Math.* **10** 29–35.

ROGERS, W. F. (1971). Exact null distributions and asymptotic expansions for rank test statistics. Technical Report No. 145, Departments of O.R. and Statistics, Stanford Univ.

STIGLER, S. M. (1969). Linear functions of order statistics. *Ann. Math. Statist.* **40** 770–788.

TITCHMARSH, E. C. (1939). *The Theory of Functions*, 2nd ed. Oxford Univ. Press.

W. ALBERS AND P. J. BICKEL
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

W. R. VAN ZWET
CENTRAAL REKEN-INSTITUUT
DER RIJKSUNIVERSITEIT TE LEIDEN
WASSENAARSEWEG 80
POSTBUS 2060
LEIDEN, THE NETHERLANDS

The Annals of Statistics
1974, Vol. 2, No. 1, 1-20

SPECIAL INVITED PAPER

EDGEWORTH EXPANSIONS IN NONPARAMETRIC STATISTICS[1]

BY P. J. BICKEL

University of California, Berkeley

This is a survey of recent work on Edgeworth expansions for (M) estimates, rank tests and some other statistics arising in nonparametric models. A Berry-Esséen theorem for U-statistics which seems to be new is also proved.

1. Introduction. During the past 25 years various procedures which are not sensitive to certain departures from normality have been evolved and investigated. The study of such methods is loosely referred to as nonparametric statistics. One broad category of such procedures is that of the distribution free tests such as the permutation t test, the rank tests of Wilcoxon, Kruskal-Wallis, Spearman and Kendall, and the omnibus tests such as the two sample Smirnov test. All of these are discussed in the monograph of Hájek and Šidák [26]. Another major category is that of the various robust estimates such as those discussed in the recent Princeton study [2].

Most of the theoretical work done on these procedures has been devoted to obtaining large sample properties by establishing first order limit theorems for the statistics on which these procedures are based. In this paper I intend to discuss what is known about higher order approximations to the distribution of these statistics. In the main I shall limit myself to discussion of results obtained since the general review paper by D. Wallace which appeared in this journal in 1958, [57].

Suppose that we are given a sequence of statistics $\{T_N\}$, $N \geqq 1$, where N usually denotes sample size. In accordance with [57] we shall say that the distribution function F_N of T_N possesses an asymptotic expansion valid to $(r + 1)$ terms if there exist functions $A_0, \cdots, A_r$ such that

$$\left|F_N(x) - A_0(x) - \sum_{j=1}^r \frac{A_j(x)}{N^{j/2}}\right| = o(N^{-r/2}) . \tag{1.1}$$

If,

$$\sup_x \left|F_N(x) - A_0(x) - \sum_{j=1}^r \frac{A_j(x)}{N^{j/2}}\right| = o(N^{-r/2}) \tag{1.2}$$

Received January 1973; revised May 1973.

[1] This research was supported by the Office of Naval Research, Contract N00014-69-A-0200-1038.

AMS 1970 *subject classifications.* Primary 62G05, 10, 20, 30, 35; Secondary 60F05.

Key words and phrases. Edgeworth, Cornish-Fisher, expansions, Berry-Esseen bounds, rank tests, (M) estimates, goodness of fit tests, U-statistics, deficiency.

we shall say the expansion is uniformly valid to $(r+1)$ terms. (This is not quite in accord with Wallace who requires the remainder to be $O(N^{-(r+1)/2})$ but is more convenient and in accord with [19].) An expansion valid to one term is just an ordinary limit theorem. It is sometimes convenient to consider expansions in which the A_j also depend on N. They are then, of course, no longer uniquely defined.

These higher order terms are of interest on various grounds.

(1) Taking one or two terms of the expansion frequently improves the basic approximation A_0 strikingly. Examples of this phenomenon may be found in Hodges and Fix [28] and Thompson, Govindarajulu and Doksum [55].

(2) The higher order terms give some qualitative insight into regions of unreliability of first order results. For instance, when the limit A_0 is normal the higher order terms A_1 and A_2 typically correct for skewness and kurtosis.

(3) The expansions can be used to discriminate between procedures equivalent to first order, as for example in Hodges and Lehmann's work on deficiency [30].

(4) Last but not least the probabilistic problems involved are very challenging.

Expansions of the type (1.1) and (1.2) are not the only ones of interest. Density functions and frequency functions of lattice random variables can sometimes be expanded. Extreme and intermediate tail probabilities can also sometimes be expanded (see for example [21], pages 517–520, [13] and [37]), and as P. Huber pointed out to me, the approximation to the power function of tests so obtained can be much more satisfactory than that based on the Edgeworth expansion. However, at least to date, the principal method used has been that of saddle point approximation which seems to require more intimate knowledge of the characteristic function of F_N than is usually available. In any case few if any such expansions appear to be available in nonparametric problems. Thus, we limit ourselves to discussion of expansions of types (1.1) ("Edgeworth") and the related expansions of F_N^{-1} ("Cornish-Fisher"). We shall deal primarily with expansions in which A_0 is the normal distribution. General results are available here for linear rank statistics (Section 2) and M estimates (Section 3) and partial results for linear combinations of order statistics and U-statistics (Section 4). What is known in nonnormal limiting situations is discussed briefly in Section 5.

2. The Berry–Esséen method and linear rank statistics. Suppose that a sequence $\{T_N\}$, $N \geqq 1$, of random variables tends to a standard normal distribution. If we let

$$\rho_N(t) = E(e^{itT_N}) \tag{2.1}$$

then we are asserting that there is a version of $\log \rho_N$ such that as $N \to \infty$,

$$\log \rho_N(t) \to -\frac{t^2}{2}\,. \tag{2.2}$$

Suppose that we have an asymptotic expansion of $\log \rho_N$ of the form,

$$(2.3) \qquad \log \rho_N(t) = -\frac{t^2}{2} + \frac{P_1(it)}{N^{\frac{1}{2}}} + \cdots + \frac{P_r(it)}{N^{r/2}} + o(N^{-r/2}),$$

where the P_j are polynomials of order $\leqq j + 2$ which vanish at 0. Such a development is plausible if the T_N have cumulants $K_{j,N}$, such that $K_{1,N} = 0$, $K_{2,N} = 1$, $K_{j,N} = O(N^{-(j-2)/2})$, $j \geqq 3$, and which themselves admit asymptotic expansions in powers of $N^{-\frac{1}{2}}$. Thus if

$$(2.4) \qquad K_{j,N} = \sum_{l=0}^{r-j+2} \frac{K_j^{(l)}}{N^{(j+l-2)/2}} + o(N^{-r/2})$$

we should have,

$$(2.5) \qquad P_k(it) = \sum_{j=3}^{k+2} \frac{K_j^{(k+2-j)}}{j!} (it)^j .$$

This is typically true although it sometimes requires a separate proof. The prototypical such T_N are, of course, standardized sums of independent identically distributed random variables. For more on expansions of the log characteristic function in terms of cumulants we refer the reader to the discussion in [57] and on pages 221–230 of [12]. Now, (2.3) corresponds to

$$(2.6) \qquad \rho_N(t) = e^{-t^2/2}\left(1 + \sum_{j=1}^{r} \frac{Q_j(it)}{N^{j/2}}\right) + o(N^{-r/2})$$

where

$$Q_1(it) = P_1(it)$$

$$Q_2(it) = P_2(it) + \frac{[P_1(it)]^2}{2}$$

and so on.

Normal Fourier inversion suggests that if

$$Q_j(it) = \sum_{k \geqq 1} a_{jk}(it)^k$$

then

$$(2.7) \qquad F_N(x) = \Phi(x) - \phi(x)\left[\sum_{j=1}^{r} \frac{1}{N^{j/2}} \sum_{k\geqq 1} a_{jk} N_{k-1}(x)\right] + o(N^{-r/2})$$

where Φ is the standard normal cdf, ϕ is the standard normal density and the N_k are Hermite polynomials defined by

$$(2.8) \qquad \frac{d^k\phi(x)}{dx^k} = (-1)^k N_k(x)\phi(x) .$$

This formal step cannot, of course, be justified in general. It fails for instance if T_N is the standardized sum of independent identically distributed lattice random variables. The passage *is* valid if the weak (2.6) can be replaced by

$$(2.9) \qquad \int_{-MN^{r/2}}^{MN^{r/2}} \left\{\left|\rho_N(t) - e^{-t^2/2}\left(1 + \sum_{j=1}^{r} \frac{Q_j(it)}{N^{j/2}}\right)\right| \Big/ |t|\right\} dt = o(N^{-r/2})$$

for every $M < \infty$. An equivalent useful form of (2.9) is

$$\int_{-\varepsilon N^{\frac{1}{2}}}^{\varepsilon N^{\frac{1}{2}}} \left\{ \left| \rho_N(t) - e^{-t^2/2}\left(1 + \sum_{j=1}^{r} \frac{Q_j(it)}{N^{j/2}}\right) \right| \Big/ |t| \right\} dt = o(N^{-r/2}) \tag{2.10}$$

and

$$\int_{\{\varepsilon N^{\frac{1}{2}} \le |t| \le MN^{r/2}\}} \frac{|\rho_N(t)|}{|t|}\, dt = o(N^{r/2})$$

for some $\varepsilon > 0$ and every $M < \infty$. That (2.9) suffices follows from a famous lemma of Berry and Esséen whose statement and proof may be found in Feller [21], Chapter 16, page 510.

The validity of (2.9) and hence of (2.7) to order $1/N$ ($r = 2$) has been established for linear rank statistics both under the hypothesis of symmetry and under contiguous location alternatives by Albers, Bickel, and van Zwet [1]. A similar expansion for the two sample Wilcoxon statistic under the null hypothesis was established earlier by Rogers [48]. Expansions for general two sample rank statistics to order $1/N$ both under the hypothesis and contiguous location alternatives are in preparation [6]. Here is a selection of the results of these papers.

Let $X_1, \cdots, X_N$ be independent identically distributed with common cdf G and density g. Let $Z_{1:N} < \cdots < Z_{N:N}$ denote the ordered $|X_j|$. Define ranks $R_1, \cdots, R_N$ by

$$|X_{R_j}| = Z_{j:N}\,.$$

Let

$$\begin{aligned} \varepsilon_j &= 1 \qquad \text{if } X_{R_j} > 0 \\ &= -1 \qquad \text{otherwise}\,, \end{aligned}$$

and suppose that $a_{1N}, \cdots, a_{NN}$ are given constants.

Define

$$T_N = \sum_{j=1}^{N} \frac{a_{jN}\,\varepsilon_j}{\sigma_N} \tag{2.11}$$

where

$$\sigma_N^2 = \sum_{j=1}^{N} a_{jN}^2\,. \tag{2.12}$$

For simplicity suppose there exists a function J on (0, 1) such that

$$a_{jN} = E(J(U_{j:N})) \tag{2.13}$$

where $U_{1:N} < \cdots < U_{N:N}$ are the order statistics of a sample of size N from the uniform distribution on (0, 1). All of the usual statistics for testing the hypothesis that g is symmetric about 0, including the sign, Wilcoxon and normal scores tests can be put in this form. Hájek and Šidák [26] provide an extensive discussion of these procedures as well as the two sample tests we shall mention.

If g is symmetric about 0 the ε_j are independent with $P[\varepsilon_j = 1] = \frac{1}{2}$. The statistic T_N is then a sum of independent nonidentically distributed random variables, and

$$\rho_N(t) = \prod_{j=1}^{N} \cos \frac{t a_{jN}}{\sigma_N}\,. \tag{2.14}$$

If $\int_0^1 J^4(t)\,dt < \infty$, Taylor expansion of (2.14) yields

$$\begin{aligned}\log \rho_N(t) &= -\frac{t^2}{2} - 2\frac{(it)^4}{4!}\sum_{j=1}^N \frac{a_{jN}^4}{\sigma_N{}^4} + o\left(\frac{1}{N}\right) \\ &= -\frac{t^2}{2} - \frac{(it)^4}{12N}\frac{\int_0^1 J^4(t)\,dt}{(\int_0^1 J^2(t)\,dt)^2} + o\left(\frac{1}{N}\right).\end{aligned} \tag{2.15}$$

If J is in addition continuously differentiable and nonconstant it is shown in [1] that (2.10) holds and hence that

$$\Phi(x) + \frac{\int_0^1 J^4(t)\,dt}{12N(\int_0^1 J^2(t)\,dt)^2}\phi(x)H_3(x)$$

is a uniformly valid expansion for F_N to three terms. In particular this proves the validity of the expansions used by Fellingham and Stoker [22] for the Wilcoxon test and by Thompson *et al.* [55] for the normal scores test up to terms of order smaller than $1/N$. Thompson *et al.* noted that the approximation using exact cumulants suggested by the first identity in (2.15) is better than the expansion suggested by the second identity while Fellingham and Stoker only considered the approximation using exact cumulants, with continuity correction. The exact cumulant Edgeworth expansion in both cases did provide substantial improvement over the normal approximation for $N = 10 - 20$ although the latter seems satisfactory for all practical purposes. It is not yet known whether the Edgeworth expansion for statistics such as the normal scores is valid to more than three terms. It seems clear that the expansion to order $1/N^2$ for the Wilcoxon with continuity correction used by Fellingham and Stoker can be justified by a local limit expansion and application of the Euler-Maclaurin formula. Local limit theorems for the two sample Wilcoxon statistic were developed by Rogers [48].

If g is not symmetric about 0 the ε_j are no longer independent. However by conditioning on $|X_1|, \cdots, |X_N|$ Albers, Bickel and van Zwet arrive at the following representation for ρ_N,

$$\rho_N(t) = E\{\textstyle\prod_{j=1}^N [P_{jN}\exp[ita_{jN}/\sigma_N] + (1 - P_{jN})\exp[-ita_{jN}/\sigma_N]]\} \tag{2.16}$$

where

$$P_{jN} = \frac{g(Z_{j:N})}{g(Z_{j:N}) + g(-Z_{j:N})}.$$

From this representation it may be shown that if $\int_0^1 J^4(t)\,dt < \infty$ and J is continuously differentiable and nonconstant then

$$\int_{-bN^{\frac{1}{2}}}^{bN^{\frac{1}{2}}} \{|\rho_N(t) - \tilde\rho_N(t)|/|t|\}\,dt \leqq cN^{-\frac{3}{2}}$$

for b, c depending on g where

$$\tilde\rho_N(t) = E\left\{\exp\left[itK_{1N} - \frac{t^2}{2}K_{2N}\right]\left(1 + \frac{(it)^3}{6}K_{3N} + \frac{(it)^4}{24}K_{4N} + \frac{(it)^6}{72}K_{3N}^2\right)\right\} \tag{2.17}$$

and

$$K_{1N} = \sum_{j=1}^{N} \frac{a_{jN}}{\sigma_N} (2P_{jN} - 1)$$

$$K_{2N} = 4 \sum_{j=1}^{N} \frac{a_{jN}^3}{\sigma_N^{\,2}} P_{jN}(1 - P_{jN})$$

$$K_{3N} = 8 \sum_{j=1}^{N} \frac{a_{jN}^3}{\sigma_N^{\,3}} P_{jN}(1 - P_{jN})(1 - 2P_{jN})$$

$$K_{4N} = 16 \sum_{j=1}^{N} \frac{a_{jN}^4}{\sigma_N^{\,4}} P_{jN}(1 - P_{jN})(1 - 6P_{jN} + 6P_{jN}^2)$$

are the cumulants of T_N.

Further expansion for fixed alternatives appears to depend on the development of the theory of Edgeworth expansion for linear combinations of order statistics. However, if we permit g to depend on N in such a way that g is contiguous to a symmetric density, then K_{1N} is to first order a constant, and further expansion is possible. Specifically suppose that

$$g_N(x) = f(x - \theta_N) \tag{2.18}$$

where f is a fixed density symmetric about 0 and $\theta_N = \theta/N^{\frac{1}{2}}$. It is then shown in [1] under some regularity conditions on f, as well as the previously specified conditions on J, that for some b, c depending on f and J

$$\int_{-bN^{\frac{3}{2}}}^{bN^{\frac{3}{2}}} \{|\tilde{\rho}_N(t) - \gamma_N(t)|/|t|\}\, dt \leqq cN^{-\frac{5}{4}} \tag{2.19}$$

where

$$\gamma_N(t) = \exp\left[it\tilde{K}_{1N} - \frac{t^2}{2}\tilde{K}_{2N}\right]\left(1 + \frac{(it)^3}{6}\tilde{K}_{3N} + \frac{(it)^4}{24}\tilde{K}_{4N}\right) \tag{2.20}$$

and

$$\tilde{K}_{1N} = -\theta_N \sum_{j=1}^{N} \frac{a_{jN}}{\sigma_N} E_0(\psi_1(Z_{j:N}))$$
$$\qquad - \frac{\theta_N^{\,3}}{3\sigma_N} \sum_{j=1}^{N} a_{jN} E_0[\tfrac{1}{2}\psi_3(Z_{j:N}) - 3\psi_1\psi_2(Z_{j:N}) + \tfrac{3}{2}\psi_1^{\,3}(Z_{j:N})]$$

$$\tilde{K}_{2N} = 1 - \theta_N^{\,2} \sum_{j=1}^{N} \frac{a_{jN}^2}{\sigma_N^{\,2}} E_0(\psi_1(Z_{j:N}))^2 + \frac{\theta_N^{\,2}}{\sigma_N^{\,2}} \operatorname{Var}_0 \left(\sum_{j=1}^{N} a_{jN}\psi_1(Z_{j:N})\right)$$

$$\tilde{K}_{3N} = 2\theta_N \sum_{j=1}^{N} \frac{a_{jN}^3}{\sigma_N^{\,3}} E_0(\psi_1(Z_{j:N}))$$

$$\tilde{K}_{4N} = -2 \sum_{j=1}^{N} \frac{a_{jN}^4}{\sigma_N^{\,4}}$$

where

$$\psi_j(x) = \frac{f^{(j)}}{f}(x)$$

and the subscript 0 indicates that calculation is carried out under f. The $\tilde{K}_{jN}$ may be shown to be the leading terms in the expansion of the cumulants of T_N

under g_N. Berry's lemma can be applied to yield as a uniformly valid expansion for $F_N(t)$ to three terms

$$\Phi(y_N) - \phi(y_N)\left\{\frac{\tilde{K}_{3N}}{6} N_2(y_N) + \frac{\tilde{K}_{4N}}{24} N_3(y_N)\right\} \tag{2.21}$$

where

$$y_N = \frac{t - \tilde{K}_{1N}}{(\tilde{K}_{2N})^{\frac{1}{2}}}.$$

This is not strictly speaking an expansion of the type we have been considering since N enters into the approximation in a complicated fashion. However, the expansion can be used in this form, for instance, to study power under normal alternatives since in this case

$$\psi_j(x) = (-1)^j H_j(x)$$

and moments of order statistics from the half normal distribution are available (cf. [34]).

If J' is defined and continuous on [0, 1] and f satisfies some mild regularity conditions, integral approximations to the $\tilde{K}_{jN}$ can be shown to hold, and a uniformly valid expansion to three terms as defined in Section 1 can be provided. This is adequate for the Wilcoxon but not the normal scores test. If we consider the distribution of the latter under normal alternatives it turns out that the $\tilde{K}_{1N}$ term does not admit an expansion of the form $A + B/N$ with A, B fixed, but rather requires a term of the form $(B \log\log N)/N$. As noted by Wallace, expansions of the type (2.7) can validly be inverted to yield expansions for percentiles (Cornish–Fisher) and hence expansions for the power functions of the rank statistics T_N. Agreement between the power function expansions for the normal scores and Wilcoxon tests obtained from (2.21) and (2.15) for normal and logistic alternatives appears to agree well with the Monte Carlo figures of Thompson *et al.* [55]. However, agreement with the Monte Carlo figures of Arnold [3] for the power function of the Wilcoxon test under Cauchy alternatives seems unsatisfactory.

In [30] Hodges and Lehmann introduced the notion of *deficiency* of a procedure with respect to an equally efficient competitor. For tests of equal level α, the deficiency is crudely defined as the limit of the difference in sample sizes required to reach equal power for the same alternative. The power functions expansions obtained in [1] are used to calculate the deficiency of the normal scores test with respect to the t test for normal alternatives. This turns out to be infinite but of the order of $\log\log N$. The results of [1] can also be used to establish that the permutation t test has deficiency 0 with respect to the t test under normal alternatives.

Suppose now that we have two samples $X_1, \cdots, X_m, Y_1, \cdots, Y_n, N = m + n$, the first sample being distributed with common density f, the second with common density g. Let $Z_{1:N} < \cdots < Z_{N:N}$ be the order statistics of the pooled sample and define

$$\begin{aligned} \varepsilon_j &= 1 \quad \text{if} \quad Z_{j:N} = Y_k \quad \text{for some} \quad k \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

A two sample linear rank statistic standardized under the null hypothesis is then given by

$$T_N = \sum_{j=1}^N a_{jN}\left(\varepsilon_j - \frac{n}{N}\right)\Big/\tau_N^2 \tag{2.22}$$

where the a_{jN} are specified scores

$$\tau_N^2 = \left[\sum_{j=1}^N (a_{jN} - \bar{a}_N)^2\right]\frac{mn}{N(N-1)} \tag{2.23}$$

and

$$\bar{a}_N = \frac{1}{N}\sum_{j=1}^N a_{jN}\,.$$

Suppose again that the a_{jN} are given by (2.13). Using a representation of the characteristic function ρ_N of T_N related to one due to Erdös and Rényi [20] and the Berry lemma, Bickel and van Zwet [6] obtain a uniformly valid expansion for the distribution function F_N of T_N to three terms if $f = g$, n/N stays bounded away from 0 and 1, $\int_0^1 J^4(t)\,dt < \infty$, and J is nonconstant and has continuous derivative. In this case,

$$F_N(x) = \Phi(x) - \phi(x)\left\{\frac{K_{3N}^*}{6}H_2(x) + \frac{K_{4N}^*}{24}H_3(x) + \frac{[K_{3N}^*]^2}{72}H_5(x)\right\} + o\left(\frac{1}{N}\right) \tag{2.24}$$

where the K_{jN}^* are the cumulants of T_N. Essentially this result was obtained by Rogers in [48] for the Wilcoxon statistic. Formal expansions were previously considered by Hodges and Fix [28]. A Berry–Esséen bound was obtained by Stoker [53]. Expansions of the power function and deficiency calculations are in progress [6]. Formal expansions of the power function were considered by Witting [58] using moment expansions due to Sundrum [54]. More Monte Carlo studies of the power functions of the two sample tests are desirable. Figures are available for the Savage test [17] when f and g are exponential densities and for the Wilcoxon and normal scores test under normal alternatives [34], [35], [41].

There are several open problems in this area. Two which I find interesting are:

(1) The extension of these results to tests of independence such as Spearman's ρ and Kendall's τ.

(2) The establishment of valid expansions for fixed alternatives.

3. Multivariate Edgeworth expansions and (M) estimates. A significant development in the theory of asymptotic expansions occurred in 1961 with the appearance of Ranga Rao's thesis on Edgeworth expansions and Berry–Esséen bounds for sums of independent random vectors. Since then there has been considerable development in the field. Some results typical of the most recent state of the art and many references to older work may be found in Bhattacharya's paper [5] in which the following theorem is announced.

Let $\{X^{(r)} = (X_1^{(r)}, \cdots, X_k^{(r)})\}$ be a sequence of independent identically distributed k dimensional random vectors. Suppose that

$$(3.1)\qquad \begin{aligned} E(X_i^{(1)}) &= 0\,, && i = 1, \cdots, k \\ E(X_i^{(1)}X_j^{(1)}) &= \delta_{ij}\,, && 1 \leqq i \leqq j \leqq k\,. \end{aligned}$$

Let

$$(3.2)\qquad \rho(u) = E(e^{iuX^{(1)}})$$

where $u = (u_1, \cdots, u_k)$ and $uX^{(1)}$ is the inner product of u and $X^{(1)}$. As usual consider the formal expansion of $\rho^N(u/N^{\frac{1}{2}})e^{|u|^2/2}$ where $|u|^2 = \sum_{i=1}^k u_i^2$, as a power series in $N^{-\frac{1}{2}}$

$$(3.3)\qquad e^{|u|^2/2}\rho^N\left(\frac{u}{N^{\frac{1}{2}}}\right) = 1 + \sum_{j=1}^{\infty} \frac{P_j(iu)}{N^{j/2}}$$

where the P_j are polynomials whose coefficients depend on the cumulants of $X^{(1)}$. Define polynomials $\tilde{P}_j$ on R^k by the property that $(2\pi)^{-k/2}e^{-|t|^2/2}\tilde{P}_j(t)$ has $e^{-|u|^2/2}P_j(iu)$ as its Fourier transform. For any $A \subset R^k$, let $(\partial A)^\varepsilon$ be the set of all points within a distance ε of the boundary of A, i.e.,

$$(3.4)\qquad (\partial A)^\varepsilon = \{x \in R^k : \exists y \in A,\, z \notin A \ni |x - y| < \varepsilon,\, |x - z| < \varepsilon\}\,.$$

Let $\mathscr{A}(\Phi : d, \varepsilon_0)$ be the class of all Borel sets A such that

$$\Phi((\partial A)^\varepsilon) \leqq d\varepsilon\,, \qquad 0 < \varepsilon \leqq \varepsilon_0$$

where Φ is the standard multivariate normal product probability measure on R^k.

We need Cramér's condition

$$(\mathrm{C})\qquad \limsup_{|u|\to\infty} |\rho(u)| < 1\,.$$

THEOREM (Remark 1, page 255 of [5]). *Suppose that* $E|X_j^{(1)}|^s < \infty$, $1 \leqq j \leqq k$, *for some* $s \geqq 3$, *the* $X^{(j)}$ *are as above and that condition* (C) *holds. Let* $S_N = \sum_{j=1}^N X^{(j)}$. *Then, for every* $d > 0$,

$$(3.5)\qquad \sup\left\{\left|P\left[\frac{S_N}{N^{\frac{1}{2}}} \in A\right] - (2\pi)^{-k/2}\int\cdots\int_A e^{-|t|^2/2} \times\left[1 + \sum_{j=1}^{s-2}\frac{\tilde{P}_j(t)}{N^{j/2}}\right]dt\right| : A \in \mathscr{A}(\Phi : d, \varepsilon_0)\right\} = o(N^{(s-2)/2})\,.$$

By making a linear transformation of the variables this result can obviously be extended to the case that $X^{(1)}$ has a specified nonsingular covariance matrix. These results have been applied in a variety of problems involving expansions of multivariate distributions connected with normal variables. An interesting paper along these lines which also faces the problem of computation of the $\tilde{P}_j(t)$ is that of Chambers [10].

In this section we review the work of Linnik and Mitrofanova [38], [56] and Čibišov [11] who employed results of this type to obtain asymptotic expansions for maximum likelihood estimates, and the related work of Pfanzagl [45], [46] and Michel and Pfanzagl [40]. The work is of interest from the point of view of robust estimation since the same technique yields expansions for Huber's (M) estimates [32], [33].

Let

$$X_j = \theta + E_j\,, \qquad 1 \leqq j \leqq N$$

where the E_j are independent identically distributed with density f. An (M) estimate (scale known) of θ, for given ψ, is by definition, any solution $\hat{\theta}$ of the equation

$$\sum_{j=1}^{N} \psi(X_j - \hat{\theta}) = 0\,. \tag{3.6}$$

For the estimation to make sense we suppose

$$E_0(\psi(X_1 - \theta)) = 0\,. \tag{3.7}$$

Condition for consistency and asymptotic normality of such estimates are given in [32] and [33].

Linnik and Mitrofanova [38], in the tradition of Cramér [12], obtained expansions for a solution of (3.6) when $\psi = -f'/f$. It is easy to see in the light of [33] how their conditions should be modified to yield expansions for (M) estimates. It should be noted that [38] has many obscure points and, in particular, it seems to me that the appeal to Ranga Rao's theorem [47] at a crucial point in [38] is inadequate. However, I believe application of the more sophisticated theorem of Bhattacharya that was stated above will carry the proof through.

The main idea which was already used by Haldane and Smith [27] and Shenton and Bowman [9] for formal cumulant expansions of maximum likelihood estimates is to expand the likelihood equation beyond the customary two terms.

$$\begin{aligned} 0 = N^{-\frac{1}{2}} \textstyle\sum_{j=1}^{N} \psi(X_j - \theta) - \left\{\frac{1}{N} \textstyle\sum_{j=1}^{N} \psi'(X_j - \theta)\right\} N^{\frac{1}{2}}(\hat{\theta} - \theta) + \cdots \\ + N^{-(k-1)/2} \frac{(-1)^k}{k!} \left\{\frac{1}{N} \textstyle\sum_{j=1}^{N} \psi^{(k)}(X_j - \theta)\right\} N^{k/2}(\hat{\theta} - \theta)^k + R_{Nk}\,. \end{aligned} \tag{3.8}$$

Using the expansion to two terms and suitable conditions on the derivatives of ψ the first step is to show that large deviations of a suitable root of (3.6) are very unlikely and hence that R_{Nk} which is governed by $N^{\frac{1}{2}}(\hat{\theta} - \theta)^{k+1}$ can be bounded by something only slightly larger than $N^{-k/2}$. The next step is to consider the equation

$$\begin{aligned} 0 = N^{-\frac{1}{2}} \textstyle\sum_{j=1}^{N} \psi(X_j - \theta) - \left\{\frac{1}{N} \textstyle\sum_{j=1}^{N} \psi'(X_j - \theta)\right\} N^{\frac{1}{2}}(t - \theta) + \cdots \\ + N^{-(k-1)/2} \frac{(-1)^k}{k!} \left\{\frac{1}{N} \textstyle\sum_{j=1}^{N} \psi^{(k)}(X_j - \theta)\right\} N^{k/2}(t - \theta)^k\,. \end{aligned} \tag{3.9}$$

The solution $t = \hat{\theta}_1$ of this equation can be expanded in an asymptotic expansion in $N^{-\frac{1}{2}}$ whose leading term is $N^{-\frac{1}{2}} \sum_{j=1}^{N} \psi(X_j - \theta)/E_\theta(\psi'(X_1 - \theta))$ and whose coefficients are polynomials in $\xi_0, \cdots, \xi_k$ where

$$\xi_r = \frac{1}{N^{\frac{1}{2}}} \sum_{j=1}^{N} [\psi^{(r)}(X_j - \theta) - E_\theta(\psi^{(r)}(X_j - \theta))]. \tag{3.10}$$

Then one shows that $\hat{\theta}$ and $\hat{\theta}_1^{(k)}$, the sum of the first k terms in the expansion of $\hat{\theta}_1$, differ to an order that matters only on a set of relatively negligible probability. Then one applies a theorem such as Bhattacharya's to the event $[N^{\frac{1}{2}}(\hat{\theta}_1^{(k)} - \theta) < x]$ which indeed depends only on $(\xi_0, \cdots, \xi_k)$. Finally there is the problem of expanding the multivariate integrals appearing in the multivariate Edgeworth theorem since these depend on N (since $\hat{\theta}_1^{(k)}$ is a polynomial in powers of $N^{-\frac{1}{2}}$ as well as in the ξ_j). The result is an expansion of the type (2.7). It is formally clear that the coefficients should agree with those obtained by using the formal expansions of the cumulants in powers of $N^{-\frac{1}{2}}$ from [27] and then proceeding to get a formal Edgeworth expansion from the formal Charlier expansion as in (2.4) and (2.5). However, this has not been checked to my knowledge.

Mitrofanova [42] extended the work of [38] to maximum likelihood estimates of a vector parameter. Unfortunately, as was noted by Pfanzagl [46], her proof contains very serious gaps. A salvage operation however seems both possible and worthwhile. In particular this should yield valid expansions for (M) estimates when scale is estimated (as it normally would be). Čibišov's announcement [11] is essentially an extension of the work of [38] to maximum likelihood estimation of a single parameter under rather simple conditions.

Pfanzagl [46] and Michel and Pfanzagl [40] have used a different approach which though much simpler for the case of a single parameter does not appear to generalize. The idea similar to that used by Huber in [32] and earlier by H. E. Daniels [14] is to compare the events $[\hat{\theta} < x]$ and $[\sum_{j=1}^{N} \psi(X_j - x) < 0]$. For increasing ψ the two events are essentially the same. In general even for functions of the form $\psi(x, \theta)$, under suitable conditions, one can argue that the difference of the two events has negligible probability for $x = \theta + a/N^{\frac{1}{2}}$ with $|a|$ bounded. But to $P[\sum_{j=1}^{N} \psi(X_j - x) < 0]$ one can apply the classical univariate expansions for sums of independent identically distributed random variables and then use suitable expansions in $(x - \theta)/N^{\frac{1}{2}}$ of the cumulants of $\psi(X_1 - x)$. This method has the advantage of enabling one to deal with ψ functions which are not very smooth such as those introduced by Huber [32]. There seems at present, however, to be no way of dealing with (M) estimates in which scale is estimated simultaneously when the functions defining the estimates cannot be expanded along the lines of [33].

Pfanzagl [46] gives a variety of applications to parametric models of the univariate expansions mentioned above. There have been hardly any numerical studies of the applicability of these expansions. An interesting example, however, is Barnett's work [4] in which he shows that the (formal) expansion is relatively poor when applied to the maximum likelihood estimate of location for a Cauchy sample.

4. Other classes of asymptotically normal statistics. There has been little success so far in validating expansions or even establishing Berry–Esséen bounds of order $1/N^{\frac{1}{2}}$ for general classes of statistics known to be asymptotically normally distributed, other than the ones we have discussed.

Mr. S. Bjerve in work towards a Berkeley thesis has shown that trimmed means admit valid Edgeworth expansions and is in the process of explicitly calculating the coefficients for comparison with the published distributions of the Princeton project [2]. His method employs special properties of the trimmed means and does not carry over to more general estimates. Further work on systematic statistics which can also be handled by elementary means is intended. Even formal work seems surprisingly scarce here. In this connection I would like to mention [16] in which expansions are obtained for the cumulants of single order statistics.

The only theoretical result on rates of convergence for general linear combinations of order statistics known to me is due to Rosenkrantz and O'Reilly [43] who establish various bounds of Berry–Esséen type for the error committed by using the normal approximation to the distribution of a linear combination of order statistics. None of these bounds is of smaller order than $N^{-\frac{1}{4}}$ where N is the sample size. This limitation appears due to the Skorokhod embedding method which they employ. This order is, of course, incorrect for all cases in which sharp bounds are available, i.e., trimmed means (including the mean) and systematic statistics. I conjecture that under mild conditions the "right" order is $N^{-\frac{1}{2}}$.

In 1948 Hoeffding [31] introduced the interesting class of U-statistics, which includes among its members the Wilcoxon two sample statistic. As another illustration of the power of the Fourier technique in a nonstandard situation we shall prove under rather strong conditions that the normal approximation to the distribution of a U-statistic of order 2 is valid to order $N^{-\frac{1}{2}}$. Our method can be adapted to yield the $N^{-\frac{1}{2}}$ bound for the one and two sample Wilcoxon statistic as well as Kendall's τ. (In fact fixed alternative asymptotic expansions for these statistics can be obtained using a combination of the methods of the appendix and those of [1].) The method should also extend to von Mises statistics [56] of order 1 and hence to linear combinations of order statistics. However we are unable to get $N^{-\frac{1}{2}}$ bounds for U-statistics with unbounded kernels. Bounds of order $N^{-r/2}$, $r < 1$, have been obtained by Grams and Serfling in [25] by a different technique. Asymptotic expansions in general seem out of reach. Here is the statement of our theorem. The proof is given in an appendix.

Let $R_1, \cdots, R_N$ be a sample from the uniform distribution on $(0, 1)$. Let ψ be a measurable real-valued function on the closed unit square such that $|\psi| \leqq M < \infty$ (say). Suppose moreover that ψ is symmetric, $\psi(u, v) = \psi(v, u)$ and that

$$\int_0^1 \int_0^1 \psi(u, v)\, du\, dv = 0\,. \tag{4.1}$$

Let

$$T_N = \frac{1}{\sigma_N} \sum_{i<j} \psi(R_i, R_j) \tag{4.2}$$

where

$$\sigma_N{}^2 = \frac{N(N-1)}{2} \int_0^1 \int_0^1 \psi^2(u, v)\, du\, dv + N(N-1)(N-2) \int_0^1 \gamma^2(u)\, du \tag{4.3}$$

and

$$\gamma(u) = \int_0^1 \psi(u, v)\, dv\,. \tag{4.4}$$

THEOREM 4.1. *If the preceding assumptions hold and γ does not vanish identically, then there exists a constant C depending on ψ but not N such that*

$$\sup_x |P[T_N \leqq x] - \Phi(x)| \leqq \frac{C}{N^{\frac{1}{2}}}$$

where Φ is the standard normal cumulative distribution function.

A new approach has recently been advanced by Stein [52] which does not rely on Fourier analytic methods. Using his method he is able to show that the error committed in applying the normal approximation to the sum of the first N of a stationary sequence of bounded m dependent random variables is of order $N^{-\frac{1}{2}}$. The possibility of applying his method to some of the classes we have considered should be investigated.

5. Expansions for statistics with nonnormal limiting distributions. The omnibus goodness of fit and two sample tests such as those of Kolmogorov–Smirnov and Cramér–von Mises and the Pearson χ^2 test do not have limiting normal distributions. The Russian school of probability theorists has had considerable success in obtaining expansions for the distribution of the Kolmogorov–Smirnov test statistics under the null hypothesis. The methods employed at first used explicit representations of the null distribution. An account of results of this type due to Chan Li-Tsien may be found in Gnedenko, Korolyuk, Skorokhod [23]. The most definitive expansion for the one-sided goodness of fit statistic was given by Lauwerier [36]. Subsequently, the problems were treated as special cases of more general problems of first passage times of random walks (cf. for example Borovkov [7] in which the two sample Smirnov statistic is treated). An account of the latest results and extensive references may be found in Borovkov [8]. Since none of the first order limiting distributions under contiguous alternatives for these statistics have been tabled or extensively studied it is not surprising that there has been no work on asymptotic expansions for the power.

There has recently been some interest in obtaining Berry–Esséen type bounds for the difference between the distribution of the Cramér–von Mises goodness of fit statistic under the null hypothesis and its well known limit distribution. However, the methods used by Rosenkrantz in [49] and Sawyer in [50] (cf. also Orlov [44]) use the Skorokhod embedding and not surprisingly obtain bounds which are of order strictly worse than $N^{-\frac{1}{2}}$ where N is the sample size. In an announcement of results without proofs [15] D. Darling obtained a representation for the characteristic function of the von Mises statistic which he employed to get an asymptotic expansion of the characteristic function to two terms for fixed argument. I do not know whether this approach can be refined to yield the kind of estimates which permit us to apply Berry's lemma.

Finally, I want to mention the recent Chicago thesis of Yarnold [59] in which he obtained asymptotic expansions for the distribution of Pearson's χ^2 statistic. Since χ^2 is a smooth function of the multinomial frequencies we might expect that the theorems on multivariate Edgeworth series should apply. Unfortunately the vector of multinomial frequencies is a normalized sum of independent identically distributed random vectors taking their values in a lattice, Cramér's condition (C) does not hold and in fact the formal Edgeworth expansion is invalid. However, it is possible to use the well-known local limit expansion for the multinomial probability and then sum up over all points in the appropriate region. This is an improvement over the χ^2 approximation but almost as complicated as calculation of the exact probabilities. Moreover, it does not yield a form which is sufficiently tractable analytically to settle long outstanding questions about the relative performance of the χ^2 and likelihood ratio tests. Results which are manageable in this area would be interesting but seem hard.

6. Appendix (Proof of Theorem 4.1). Let

$$S_N = \frac{(N-1)}{\sigma_N} \sum_{i=1}^N \gamma(R_i) \tag{6.1}$$

$$\Delta_N = T_N - S_N \tag{6.2}$$

$$\phi_N(t) = E(e^{itT_N}) \tag{6.3}$$

$$\gamma(t) = E(e^{it\gamma(R_1)}) \tag{6.4}$$

$$\tilde{\phi}_N(t) = E(e^{itS_N}) = \gamma^N\left(\frac{t(N-1)}{\sigma_N}\right). \tag{6.5}$$

The crux of the argument is to show that there exists $\varepsilon_1 > 0$ and a constant D_1 both independent of N such that

$$\int_{-\varepsilon_1 N^{\frac{1}{2}}}^{\varepsilon_1 N^{\frac{1}{2}}} \frac{|\phi_N(t) - \tilde{\phi}_N(t)|}{|t|}\, dt \leqq D_1 N^{-\frac{1}{2}}. \tag{6.6}$$

Since it is well known that there exists $\varepsilon_2 > 0$ and a constant D_2 both independent of N such that

$$\int_{-\varepsilon_2 N^{\frac{1}{2}}}^{\varepsilon_2 N^{\frac{1}{2}}} \frac{|\tilde{\phi}_N(t) - e^{-t^2/2}|}{|t|}\, dt \leqq D_2 N^{-\frac{1}{2}},$$

it follows that if $\varepsilon = \min(\varepsilon_1, \varepsilon_2)$, $D = D_1 + D_2$,

$$\int_{-\varepsilon N^{\frac{1}{2}}}^{\varepsilon N^{\frac{1}{2}}} \frac{|\phi_N(t) - e^{-t^2/2}|}{|t|}\, dt \leqq D N^{-\frac{1}{2}}, \tag{6.7}$$

and the theorem follows from (6.6) and the usual Berry–Esséen argument.

To prove (6.6) we need the following lemmas.

LEMMA 6.1. *Let* $\{\xi_j\}$, $1 \leqq j \leqq n$ *be a sequence of martingale summands, i.e.,*

$$E(\xi_j \mid \xi_1, \cdots, \xi_{j-1}) = 0\,, \qquad 1 \leqq j \leqq n\,.$$

Let $W_n = \sum_{j=1}^n \xi_j$. *Define* $m_{n,k} = \max_{1 \leq j \leq n} E(\xi_j^{2k})$, $k \geqq 1$. *Then, for* $k \leqq n$,

$$E(W_n^{2k}) \leqq n^k m_{n,k}(4ek)^k\,. \tag{6.8}$$

REMARKS. (1) An estimate similar to (6.8) has been obtained by Dharmadhikari, Fabian and Jogdeo [18] with $m_{n,k}$ replaced by $(1/n)\sum_{j=1}^n E(\xi_j^{2k})$. However, their bound grows with k as 2^{k^2} which is quite inadequate for our purposes. We note that our technique readily establishes,

$$E(W_n^{2k}) \leqq n^k m_{n,k}(k)^{2k}$$

for all k, n but even this is inadequate.

(2) The example of ξ_j i.i.d. normal random variables with mean 0 shows that our bound is comparatively sharp. Also see the remark on Lemma 6.2.

Our main interest in Lemma 6.1 is in its application to

LEMMA 6.2. *Under the conditions of Theorem* 4.1, *if* $k \leqq N$,

$$E(\Delta_N^{2k}) \leqq \sigma_N^{-2k} N^{2k}(3M)^{2k}(4ek)^{2k}\,. \tag{6.9}$$

REMARK. The order of magnitude of the coefficient of $\sigma^{N-2k}N^{2k}$ in (6.9) is quite sharp. Thus if $\psi(x, y) = \frac{3}{4}$ if x and y are both $\geqq \frac{1}{2}$, $= -\frac{1}{4}$ otherwise

$$\sigma_N \Delta_N = \sum_{i<j} \eta_i \eta_j = \frac{1}{2}\left[(\textstyle\sum_{i=1}^N \eta_i)^2 - \dfrac{N}{4}\right] \tag{6.10}$$

where the η_i are independent and equal $\pm\frac{1}{2}$ with equal probability $\frac{1}{2}$. It is easy to see that

$$E(\sigma_N \Delta_N)^{2k} \geqq 8^{-2k}\{2^{-2k+1}E(U_N^{4k}) - N^{2k}\} \tag{6.11}$$

where $U_N = \sum_{i=1}^N \varepsilon_i$ and $\varepsilon_i = \pm 1$ with probability $\frac{1}{2}$. Since,

$$E(U_N^{4k}) = \sum_{t_1 + \cdots + t_N = 2k} \frac{4k!}{2t_1! \cdots 2t_N!}\,,$$

$$E(U_N^{4k}) \geqq \binom{N}{2k}\frac{4k!}{2^{2k}} \geqq A(kN)^{2k}\left(1 - \frac{(2k-1)}{N}\right)^{2k}\left(\frac{4}{e}\right)^{2k}$$

for some universal constant A and hence,

$$(kN)^{-2k}E(\sigma_N \Delta_N)^{2k} \geqq c\rho^k$$

for all N and $k \leqq aN$, $a < \frac{1}{2}$ where c and ρ depend on a but not on k and N. Then the ratio between $E(\sigma_N \Delta_N)^{2k}$ and the estimate given by (6.9) is (relatively) negligible.

PROOF OF LEMMA 6.1. The proof is by induction on n for fixed k. Note first that

$$E(\xi_1 + \cdots + \xi_k)^{2k} \leqq k^{2k} m_{k,k} \tag{6.12}$$

and hence the induction hypothesis holds for $n = k$. Suppose it is true for $n = l \geqq k$. Then

$$E(W_{l+1}^{2k}) = E(W_l^{2k}) + \textstyle\sum_{j=2}^{2k} \binom{2k}{j} E(W_l^{2k-j}\xi_{l+1}^j) \tag{6.13}$$

by the martingale hypothesis. By induction and the Hölder inequality we obtain

$$\begin{aligned} E(W_l^{2k-j}\xi_{l+1}^j) &\leqq [c_k l^k m_{l,k}]^{1-j/2k}[m_{l+1,k}]^{j/2k} \\ &\leqq (c_k l^k m_{l+1,k})(c_k^{1/2k} l^{\frac{1}{2}})^{-j} \end{aligned} \tag{6.14}$$

where $c_k = (4ek)^k$. By elementary estimates (6.13) and (6.14) yield

$$\begin{aligned} E(W_{l+1}^{2k}) &\leqq c_k l^k m_{l+1,k} \left(1 + \frac{4k^2}{l c_k^{1/k}} \textstyle\sum_{j=0}^{2k-2} \binom{2k-2}{j} (c_k^{1/2k} l^{\frac{1}{2}})^{-j}\right) \\ &\leqq c_k l^k m_{l+1,k} \left(1 + \frac{k}{le}\left(1 + \frac{1}{2(ekl)^{\frac{1}{2}}}\right)^{2k-2}\right) \\ &\leqq c_k l^k m_{l+1,k} \left(1 + \frac{k}{l}\right) \end{aligned} \tag{6.15}$$

for $k \leqq l$. Since $(1 + k/l) \leqq ((l+1)/l)^k$ the hypothesis is verified for $n = l + 1$ and the result follows.

Proof of Lemma 6.2. Begin by noting that

$$\sigma_N \Delta_N = \textstyle\sum_{j=1}^N \xi_j \tag{6.16}$$

where

$$\xi_j = \textstyle\sum_{i=1}^{j-1} [\psi(R_i, R_j) - \gamma(R_i) - \gamma(R_j)] \tag{6.17}$$

and that the ξ_j are martingale summands. Moreover, note that

$$E(\xi_j^{2k}) = E(E[\textstyle\sum_{i=1}^{j-1} (\psi(R_i, R_j) - \gamma(R_i) - \gamma(R_j))^{2k} \mid R_j]) \tag{6.18}$$

and that given R_j the summands $\eta_i = (\psi(R_i, R_j) - \gamma(R_i) - \gamma(R_j))$, $i = 1, \cdots, j-1$ are also martingale summands (in fact i.i.d.). Since

$$E(\psi(R_1, R_2) - \gamma(R_1) - \gamma(R_2))^{2k} \leqq (3M)^{2k} \tag{6.19}$$

we can apply Lemma 6.1 twice in succession to obtain Lemma 6.2.

Lemma 6.3. *Under the conditions of the theorem,*

$$|E(e^{itS_N}\Delta_N)| \leqq 3M^3 t^2 \frac{N^4}{\sigma_N^3} |\eta|^{N-2}\left(\frac{t}{\sigma_N}(N-1)\right) \tag{6.20}$$

$$|E(e^{itS_N}\Delta_N^j)| \leqq \left(\frac{N^2}{\sigma_N}\right)^j \left(\frac{3M}{2}\right)^j |\eta|^{N-2j}\left((N-1)\frac{t}{\sigma_N}\right) \quad \textit{for } j \geqq 1\,. \tag{6.21}$$

Proof. To prove (6.20) we calculate

$$\begin{aligned} E(\Delta_N e^{itS_N}) = {} & \frac{N(N-1)}{2\sigma_N} \eta^{N-2}\left(\frac{t}{\sigma_N}(N-1)\right) \\ & \times E\left(\exp\left[\frac{it(N-1)}{\sigma_N}(\gamma(R_1) + \gamma(R_2))\right]\right. \\ & \left.\times (\psi(R_1, R_2) - \gamma(R_1) - \gamma(R_2))\right). \end{aligned} \tag{6.22}$$

Since $\psi(R_1, R_2) - \gamma(R_1) - \gamma(R_2)$ and $\gamma(R_1), \gamma(R_2)$ are uncorrelated we can write

$$\begin{aligned} &\left|E\left(\exp\left[\frac{it(N-1)}{\sigma_N}(\gamma(R_1)+\gamma(R_2))\right](\psi(R_1, R_2) - \gamma(R_1) - \gamma(R_2))\right)\right| \\ &\quad = \left|E\left[\left(\exp\left[\frac{it(N-1)}{\sigma_N}(\gamma(R_1)+\gamma(R_2))\right] - 1\right)\right.\right. \\ &\qquad \left.\left.\times (\psi(R_1, R_2) - \gamma(R_1) - \gamma(R_2))\right]\right| \\ &\quad \leqq \frac{t^2}{2}\frac{(N-1)^2}{{\sigma_N}^2} E[(\gamma(R_1)+\gamma(R_2))^2|\psi(R_1, R_2) - \gamma(R_1) - \gamma(R_2)|] \\ &\quad \leqq 6M^3t^2\frac{(N-1)^2}{{\sigma_N}^2}, \end{aligned} \tag{6.23}$$

and (6.20) follows.

Similarly,

$$\begin{aligned} &{\sigma_N}^j E({\Delta_N}^j e^{itS_N}) \\ &\quad = \sum_{\{(a_1,b_1),\cdots,(a_j,b_j)\}} E(e^{itS_N}[\textstyle\prod_{i=1}^j (\psi(R_{a_i}, R_{b_i}) - \gamma(R_{a_i}) - \gamma(R_{b_i}))]) . \end{aligned} \tag{6.24}$$

Applying elementary inequalities we obtain

$$\begin{aligned} &|{\sigma_N}^j E({\Delta_N}^j e^{itS_N})| \\ &\quad \leqq \frac{N^{2j}}{2^j}|\eta|^{N-2j}\left((N-1)\frac{t}{\sigma_N}\right) E|\psi((R_1, R_2) - \gamma(R_1) - \gamma(R_2))|^j \\ &\quad \leqq \left(\frac{3M}{2}\right)^j N^{2j}|\eta|^{N-2j}\left(\frac{(N-1)t}{\sigma^N}\right). \end{aligned} \tag{6.25}$$

The lemma follows.

We proceed with the proof of (6.6). Since

$$|\phi_N(t) - \tilde{\phi}_N(t)| = |E(e^{itS_N}(e^{it\Delta_N} - 1))|$$

we have for any k,

$$|\phi_N(t) - \tilde{\phi}_N(t)| \leqq \left|\textstyle\sum_{j=1}^{2k-1}\dfrac{(it)^j}{j!}E(e^{itS_N}{\Delta_N}^j)\right| + \frac{t^{2k}}{(2k)!}E({\Delta_N}^{2k}) . \tag{6.26}$$

From (6.26), (6.9) and (6.20),

$$|\phi_N(t) - \tilde{\phi}_N(t)| \leqq \left(3M^3\frac{N^4}{{\sigma_N}^3}|\eta|^{N-2}\left(\frac{t}{\sigma_N}(N-1)\right)t + 8e^2\frac{N^2}{{\sigma_N}^2}M^2\right)t^2 . \tag{6.27}$$

Since there exists $\theta > 0$ such that ${\sigma_N}^2 \geqq \theta^2 N^3$ for all N we conclude that

$$\begin{aligned} &\int_{-N^{\frac{1}{2}}}^{N^{\frac{1}{2}}}\frac{|\phi_N(t) - \tilde{\phi}_N(t)|}{|t|}dt \\ &\quad \leqq \frac{3N^{-\frac{1}{2}}M^3}{\theta^3}\int_{-N^{\frac{1}{2}}}^{N^{\frac{1}{2}}}|t|^2|\eta|^{N-2}\left(\frac{tN^{-\frac{1}{2}}}{\theta}\right)dt + \frac{8e^2M^2}{\theta^2}N^{-\frac{1}{2}} \\ &\quad \leqq FN^{-\frac{1}{2}} \end{aligned} \tag{6.28}$$

where F is a constant depending on ψ but not N.

Let

$$\varepsilon = \frac{\theta p}{24Me}, \qquad p < 1 \tag{6.29}$$

$$k = \left\{\left(\left[\frac{1}{2}\frac{\log N}{|\log p|}\right] + 1\right) \wedge N\right\}.$$

If $|t| \leqq \varepsilon N^{\frac{1}{2}}$, by Lemma 6.2 for this k and N sufficiently large,

$$\begin{aligned}\frac{t^{2k}}{(2k)!} E(\Delta_N^{2k}) &\leqq \frac{\varepsilon^{2k} N^k}{(2k)!} \cdot \frac{N^{2k}}{\sigma_N^{2k}} k^{2k}(12eM)^{2k} \\ &\leqq \binom{4k}{2k} 2^{-4k} p^{2k} < N^{-1}.\end{aligned} \tag{6.30}$$

To complete the argument note that for p sufficiently small, there exists $\tau > 0$ such that for $|t| \leqq \varepsilon N^{\frac{1}{2}}$,

$$\log|\gamma|\left(\frac{(N-1)t}{\sigma_N}\right) \leqq -\frac{\tau t^2}{N}. \tag{6.31}$$

Applying (6.31) and (6.21) we conclude that for $N^{\frac{1}{2}} \leqq |t| \leqq \varepsilon N^{\frac{1}{2}}$, $j < 2k$,

$$|E(e^{itS_N}\Delta_N{}^j)| \leqq \theta^{-j} N^{j/2} \left(\frac{3M}{2}\right)^j \exp\left[-\tau N^{\frac{1}{2}}\left(1 - \frac{4k}{N}\right)\right]. \tag{6.32}$$

Hence for $N^{\frac{1}{2}} \leqq |t| \leqq \varepsilon N^{\frac{1}{2}}$ with k, ε given by (6.29),

$$\begin{aligned}\left|\sum_{j=1}^{2k-1} \frac{(it)^j}{j!} E(e^{itS_N}\Delta_N{}^j)\right| &\leqq eN^{2k}\left(\frac{3M}{2\theta}\right)^{2k} \exp\left[-\tau N^{\frac{1}{2}}\left(1 - \frac{4k}{N}\right)\right] \\ &= o\left(\frac{1}{N}\right)\end{aligned} \tag{6.33}$$

uniformly for $|t|$ as above. Combining (6.28), (6.30) and (6.33), (6.6) and the theorem follows.

REFERENCES

[1] ALBERS, W., BICKEL, P. J. and VAN ZWET, W. R. (1972). Asymptotic expansions for the power of distribution free tests in the one sample problem. To be submitted to *Ann. Statist.*

[2] ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., ROGERS, W. H. and TUKEY, J. W. (1972). *Robust Estimation: Survey and Advances.* Princeton Univ. Press.

[3] ARNOLD, H. J. (1965). Small sample power for the one sample Wilcoxon test. *Ann. Math. Statist.* **36** 1767–1778.

[4] BARNETT, V. (1966). Evaluation of maximum likelihood estimate when likelihood has multiple roots. *Biometrika* **53** 151–165.

[5] BHATTACHARYA, R. N. (1971). Rates of convergence and asymptotic expansions. *Ann. Math. Statist.* **42** 241–259.

[6] BICKEL, P. J. and VAN ZWET, W. R. (1973). Asymptotic expansions for the power of distribution free tests in the two sample problem. In preparation.

[7] BOROVKOV, A. A. (1962). On the problem of two samples. *Selected Transl. Math. Statist. Prob.* **5** 285–307.

[8] BOROVKOV, A. A. (1970). Limit theorems for random walks with boundaries. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* **3** 19–30.

[9] Bowman, K. and Shenton, L. R. (1963). Higher moments of maximum likelihood estimates. *J. Roy. Statist. Soc. Ser. B* **25** 305-317.

[10] Chambers, J. M. (1967). On methods of asymptotic approximation for multivariate distributions. *Biometrika* **54** 367-384.

[11] Čibišov, D. M. (1972). Asymptotic expansions for maximum likelihood estimates. *Teor. Verojatnost. i Primenen.* **17** 387-388.

[12] Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton Univ. Press.

[13] Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.* **25** 631-650.

[14] Daniels, H. E. (1960). The asymptotic efficiency of a maximum likelihood estimator. *Proc. Fourth Berkeley Symp. Math. Statist.* **1** 151-164.

[15] Darling, D. (1960). Sur Les Théorèmes de Kolmogorov-Smirnov. *Theor. Probability Appl.* **5** 356-361.

[16] David, F. N. and Johnson, N. L. (1954). Statistical treatment of censored data, Part I. *Biometrika* **41** 228-240.

[17] Davies, R. B. (1971). Rank tests for Lehmann's alternative. *J. Amer. Statist. Assoc.* **66** 879-883.

[18] Dharmadhikari, S. W., Fabian, V. and Jogdeo, K. (1968). Bounds on the moments of martingales. *Ann. Math. Statist.* **39** 1719-1723.

[19] Erdelyi, A. (1956). *Asymptotic Expansions.* Dover, New York.

[20] Erdös, P. and Rényi, A. (1950). On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hungar. Acad. Sci.* **4** 49-61.

[21] Feller, W. (1966). *An Introduction to Probability Theory and its Applications*, **2.** Wiley, New York.

[22] Fellingham, S. A. and Stoker, D. J. (1964). An approximation for the exact distribution of Wilcoxon test for symmetry. *J. Amer. Statist. Assoc.* **59** 899-905.

[23] Gnedenko, B. V., Korolyuk, V. and Skorokhod, A. V. (1960). Asymptotic expansions in probabilty theory. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **2** 153-169.

[24] Govindarajulu, Z. and Haynam, G. (1956) Exact power of Mann-Whitney test for exponential and rectangular alternatives. *Ann. Math. Statist.* **37** 945-953.

[25] Grams, W. F. and Serfling, R. J. (1973). Convergence rates for *U*-statistics and related statistics. *Ann. Statist.* **1** 153-160.

[26] Hájek, J. and Šidák, A. (1967). *Theory of Rank Tests.* Academic Press, New York.

[27] Haldane, J. B. S. and Smith, S. M. (1956). Sampling distribution of a maximum likelihood estimate. *Biometrika* **43** 96-103.

[28] Hodges, J. L. and Fix, E. (1955). Significance probabilities of the Wilcoxon test. *Ann. Math. Statist.* **26** 301-312.

[29] Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34** 598-611.

[30] Hodges, J. L. and Lehmann, E. L. (1970). Deficiency. *Ann. Math. Statist.* **41** 783-801.

[31] Hoeffding, W. (1948). A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.* **19** 293-325.

[32] Huber, P. J. (1964). Robust estimation of location. *Ann. Math. Statist.* **35** 73-101.

[33] Huber, P. J. (1965). Behavior of maximum likelihood estimates under non-standard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1** 221-233.

[34] Klotz, J. (1963). Small sample power and efficiency of the one sample Wilcoxon and normal scores tests. *Ann. Math. Statist.* **34** 624-632.

[35] Klotz, J. (1964). On the normal scores two sample rank test. *J. Amer. Statist. Assoc.* **59** 652-654.

[36] Lauwerier, H. A. (1963). The asymptotic expansions of the statistical distribution of N. V. Smirnov. *Z. Warscheinlichkeitstheorie und Verw. Gebiete* **2** 61-68.

[37] Linnik, J. (1960). On the probability of large deviations for the sums of independent random variables. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **2** 289-306.

[38] LINNIK, J. and MITROFANOVA, N. (1965). Some asymptotic expansions for maximum likelihood estimates. *Sankhyā A* **27** 73–82.

[39] MAAG, U. and DICAIRE, G. (1971). On Kolmogorov-Smirnov type one sample statistics. *Biometrika* **58** 653–656.

[40] MICHEL, R. and PFANZAGL, J. (1971). Accuracy of normal approximations for minimum contrast estimates. *Z. Warscheinlichkeitstheorie und Verw. Gebiete* **18** 73–84.

[41] MILTON, R. C. (1970). *Rank Order Probabilities.* Wiley, New York.

[42] MITROFANOVA, N. (1967). Asymptotic expansions for maximum likelihood estimates of a vector parameter. *Theor. Probability Appl.* **12** 364–372.

[43] O'REILLY, N. and ROSENKRANTZ, W. (1972). Applications of the Skorokhod representation to rates of convergence for linear combinations of order statistics. *Ann. Math. Statist.* **43** 1204–1212.

[44] ORLOV, A. I. (1971). Estimates of speed of convergence to their limit distributions of certain statistics. *Theor. Probability Appl.* **16** 526.

[45] PFANZAGL, J. (1971) Berry-Esséen bound for minimum contrast estimates. *Metrika* **17** 82–91.

[46] PFANZAGL, J. (1973). Asymptotic expansions related to minimum contrast estimators. *Ann. Statist.* **1** 993–1026.

[47] RANGA RAO, R. (1961). On the central limit theorem in R^k. *Bull. Amer. Math. Soc.* **67** 359–361.

[48] ROGERS, W. F. (1971). Exact null distributions and asymptotic expansions for rank test statistics. Technical Report, Stanford Univ.

[49] ROSENKRANTZ, W. (1969). A rate of convergence for the von Mises functional. *Trans. Amer. Math. Soc.* **139** 329–337.

[50] SAWYER, S. (1972). Rates of convergence for some functionals in probability. *Ann. Math. Statist.* **43** 273–284.

[51] SAZONOV, V. (1969). Improvement of a convergence rate estimate. *Theor. Probability Appl.* **14** 649–651.

[52] STEIN, C. (1970). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* **2**.

[53] STOKER, D. J. (1954). An upper bound for the deviation between the distribution of Wilcoxon's test statistic for the two sample problem and its limiting normal distribution for finite samples, I, II. *Indag. Math.* **16** 599–606, 607–614.

[54] SUNDRUM, R. (1954). A further approximation to the distribution of Wilcoxon's statistic. *J. Roy. Statist. Soc. Ser. B* **16** 266–268.

[55] THOMPSON, R., GOVINDARAJULU, Z. and DOKSUM, K. (1967). Distribution and power of the absolute normal scores test. *J. Amer. Statist. Assoc.* **62** 966–975.

[56] VON MISES, R. (1947). Differentiable statistical functions. *Ann. Math. Statist.* **18** 309–348.

[57] WALLACE, D. (1958). Asymptotic approximations to distributions. *Ann. Math. Statist.* **29** 635–654.

[58] WITTING, H. (1960). A generalized Pitman efficiency. *Ann. Math. Statist.* **31** 405–414.

[59] YARNOLD, J. (1968). The accuracy of seven approximations for the null distribution of the chi-square goodness of fit statistic. Thesis, Univ. of Chicago.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

Chapter 2
Robust Statistics

Peter Bühlmann

2.1 Introduction to Three Papers on Robustness

2.1.1 General Introduction

This is a short introduction to three papers on robustness, published by Peter Bickel as single author in the period 1975–1984: "One-step Huber estimates in the linear model" (Bickel 1975), "Parametric robustness: small biases can be worthwhile" (Bickel 1984a), and "Robust regression based on infinitesimal neighbourhoods" (Bickel 1984b). It was the time when fundamental developments and understanding in robustness took place, and Peter Bickel has made deep contributions in this area. I am trying to place the results of the three papers in a new context of contemporary statistics.

2.1.2 One-Step Huber Estimates in the Linear Model

The paper by Bickel (1975) about the following procedure. Given a $\sqrt{n}$-consistent initial estimator $\tilde{\theta}$ for an unknown parameter θ, performing one Gauss-Newton iteration with respect to the objective function to be optimized leads to an asymptotically efficient estimator. Interestingly, this results holds even when the MLE is not efficient, and it is equivalent to the MLE if the latter is efficient. Such a result was known for the case where the loss function corresponds to the maximum likelihood estimator (Le Cam 1956). Bickel (1975) extends this result to much more general loss functions and models.

P. Bühlmann (✉)
ETH Zürich, Rämistrasse 101, HG G17 8092, Zürich, Switzerland
e-mail: buhlmann@stat.math.ethz.ch

J. Fan et al. (eds.), *Selected Works of Peter J. Bickel*, Selected Works in Probability and Statistics 13, DOI 10.1007/978-1-4614-5544-8_2,

The idea of a computational short-cut without sacrificing statistical was relevant more than 30 years ago (summary point 5 in Sect. 3 of Bickel 1975). Yet, the idea is still very important in large scale and high-dimensional applications nowadays. Two issues emerge.

In some large-scale problems, one is willing to pay a price in terms of statistical accuracy while gaining substantially with respect to computing power. Peter Bickel has recently co-authored a paper on this subject (Meinshausen et al. 2009): having some sort of guarantee on statistical accuracy is then highly desirable. Results as in Bickel (1975), probably of weaker form which do not touch on the concept of efficiency, are underdeveloped for large-scale problems.

The other issue concerns the fact that iterations in algorithms correspond to some form of (algorithmic) regularization which is often very effective for large datasets. A prominent example of this is with boosting: instead of a Gauss-Newton step, boosting proceeds with Gauss-Southwell iterations which are coordinatewise up-dates based on an n-dimensional approximate gradient vector (where n denotes sample size). It is known, at least for some cases, that boosting with such Gauss-Southwell iterations achieves minimax convergence rate optimality (Bissantz et al. 2007; Bühlmann and Yu 2003) while being computationally attractive. Furthermore, in view of robustness, boosting can be easily modified such that each Gauss-Southwell up-date is performed in a robust way and hence, the overall procedure has desirable robustness properties (Lutz et al. 2008). As discussed in Sect. 3 of Bickel (1975), the starting value (i.e., the initial estimator) matters also in robustified boosting.

2.1.3 Parametric Robustness: Small Biases Can Be Worthwhile

The following problem is studied in Bickel (1984a): construct an estimator that performs well for a particular parametric model $\mathscr{M}_0$ while its risk is upper-bounded for another larger parametric model $\mathscr{M}_1 \supset \mathscr{M}_0$. As an interpretation, one believes that $\mathscr{M}_0$ is adequate but one wants to guard against deviations coming from $\mathscr{M}_1$. It is shown in the paper that the corresponding optimality problem has not an explicit solution: however, approximate answers are presented and interesting connections are developed to the Efron-Morris (Efron and Morris 1971) family of translation estimates, i.e., adding a soft-thresholded additional correction term to the optimal estimator under $\mathscr{M}_0$. (The reference Efron and Morris (1971) is appearing in the text but is missing in the list of references in Bickel's paper).

The notion of parametric robustness could be interesting in high-dimensional problems. Guarding against specific deviations (which may be easier to specify in some applications than in others) can be more powerful than trying to protect nonparametrically against point-mass distributions in any direction. In this sense, this paper is a key reference for developing effective high-dimensional robust inference.

2.1.4 *Robust Regression Based on Infinitesimal Neighbourhoods*

Robust regression is analyzed in Bickel (1984b) using a nice mathematical framework where the perturbation is within a $1/\sqrt{n}$-neighbourhood of the uncontaminated ideal model. The presented results in Bickel (1984b) give a clear (mathematical) interpretation of various procedures and suggest new robust methods for regression.

A major issue in robust regression is to guard against contaminations in X-space. Bickel (1984b) gives nice insights for the classical case where the dimension of X is relatively small: a new challenge is to deal with robustness in high-dimensional regression problems where the dimension of X can be much larger than sample size. One attempt has been to robustify high-dimensional estimators such as the Lasso (Khan et al. 2007) or L_2Boosting (Lutz et al. 2008), in particular with respect to contaminations in X-space. An interesting and different path has been initiated by Friedman (2001) with tree-based procedures which are robust in X-space (in connection with a robust loss function for the error). There is clearly a need of a unifying theory, in the spirit of Bickel (1984b), for robust regression when the dimension of X is large.

References

Begun JM, Hall WJ, Huang W-M, Wellner JA (1983) Information and asymptotic efficiency in parametric–nonparametric models. Ann Stat 11(2):432–452

Beran R (1974) Asymptotically efficient adaptive rank estimates in location models. Ann Stat 2:63–74

Bickel P (1975) One-step Huber estimates in the linear model. J Am Stat Assoc 70:428–434

Bickel PJ (1982) On adaptive estimation. Ann Stat 10(3):647–671

Bickel P (1984a) Parametric robustness: small biases can be worthwhile. Ann Stat 12:864–879

Bickel P (1984b) Robust regression based on infinitesimal neighbourhoods. Ann Stat 12:1349–1368

Bickel PJ, Klaassen CAJ (1986) Empirical Bayes estimation in functional and structural models, and uniformly adaptive estimation of location. Adv Appl Math 7(1):55–69

Bickel PJ, Ritov Y (1987) Efficient estimation in the errors in variables model. Ann Stat 15(2):513–540

Bickel PJ, Ritov Y (1988) Estimating integrated squared density derivatives: sharp best order of convergence estimates. Sankhyā Ser A 50(3):381–393

Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA (1993) Efficient and adaptive estimation for semiparametric models. Johns Hopkins series in the mathematical sciences. Johns Hopkins University Press, Baltimore

Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA (1998) Efficient and adaptive estimation for semiparametric models. Springer, New York. Reprint of the 1993 original

Birgé L, Massart P (1993) Rates of convergence for minimum contrast estimators. Probab Theory Relat Fields 97(1–2):113–150

Birgé L, Massart P (1995) Estimation of integral functionals of a density. Ann Stat 23(1):11–29

Bissantz N, Hohage T, Munk A, Ruymgaart F (2007) Convergence rates of general regularization methods for statistical inverse problems and applications. SIAM J Numer Anal 45:2610–2636
Bühlmann P, Yu B (2003) Boosting with the L_2 loss: regression and classification. J Am Stat Assoc 98:324–339
Efron B (1977) The efficiency of Cox's likelihood function for censored data. J Am Stat Assoc 72(359):557–565
Efron B, Morris C (1971) Limiting the risk of Bayes and empirical Bayes estimators – part I: Bayes case. J Am Stat Assoc 66:807–815
Friedman J (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232
Hájek J (1962) Asymptotically most powerful rank-order tests. Ann Math Stat 33:1124–1147
Khan J, Van Aelst S, Zamar R (2007) Robust linear model selection based on least angle regression. J Am Stat Assoc 102:1289–1299
Kiefer J, Wolfowitz J (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann Math Stat 27:887–906
Klaassen CAJ (1987) Consistent estimation of the influence function of locally asymptotically linear estimators. Ann Stat 15(4):1548–1562
Kosorok MR (2009) What's so special about semiparametric methods? Sankhyā 71(2, Ser A): 331–353
Laurent B, Massart P (2000) Adaptive estimation of a quadratic functional by model selection. Ann Stat 28(5):1302–1338
Le Cam L (1956) On the asymptotic theory of estimation and testing hypotheses. In: Proceedings of the third Berkeley symposium on mathematical statistics and probability, vol 1. University of California Press, Berkeley, pp 129–156
Lutz R, Kalisch M, Bühlmann P (2008) Robustified L2 boosting. Comput Stat Data Anal 52:3331–3341
Meinshausen N, Bickel P, Rice J (2009) Efficient blind search: optimal power of detection under computational cost constraint. Ann Appl Stat 3:38–60
Murphy SA, van der Vaart AW (1996) Likelihood inference in the errors-in-variables model. J Multivar Anal 59(1):81–108
Neyman J, Scott EL (1948) Consistent estimates based on partially consistent observations. Econometrica 16:1–32
Pfanzagl J (1990a) Estimation in semiparametric models. Lecture notes in statistics, vol 63. Springer, New York. Some recent developments
Pfanzagl J (1990b) Large deviation probabilities for certain nonparametric maximum likelihood estimators. Ann Stat 18(4):1868–1877
Pfanzagl J (1993) Incidental versus random nuisance parameters. Ann Stat 21(4):1663–1691
Reiersol O (1950) Identifiability of a linear relation between variables which are subject to error. Econometrica 18:375–389
Ritov Y, Bickel PJ (1990) Achieving information bounds in non and semiparametric models. Ann Stat 18(2):925–938
Robins J, Tchetgen Tchetgen E, Li L, van der Vaart A (2009) Semiparametric minimax rates. Electron J Stat 3:1305–1321
Schick A (1986) On asymptotically efficient estimation in semiparametric models. Ann Stat 14(3):1139–1151
Stein C (1956) Efficient nonparametric testing and estimation. In: Proceedings of the third Berkeley symposium on mathematical statistics and probability 1954–1955, vol. I. University of California Press, Berkeley/Los Angeles, pp 187–195
Stone CJ (1975) Adaptive maximum likelihood estimators of a location parameter. Ann Stat 3:267–284
Strasser H (1996) Asymptotic efficiency of estimates for models with incidental nuisance parameters. Ann Stat 24(2):879–901
Tchetgen E, Li L, Robins J, van der Vaart A (2008) Minimax estimation of the integral of a power of a density. Stat Probab Lett 78(18):3307–3311

van der Vaart AW (1988) Estimating a real parameter in a class of semiparametric models. Ann Stat 16(4):1450–1474

van der Vaart A (1991) On differentiable functionals. Ann Stat 19(1):178–204

van der Vaart A (1996) Efficient maximum likelihood estimation in semiparametric mixture models. Ann Stat 24(2):862–878

van Eeden C (1970) Efficiency-robust estimation of location. Ann Math Stat 41:172–181

Wellner JA, Klaassen CAJ, Ritov Y (2006) Semiparametric models: a review of progress since BKRW (1993). In: Frontiers in statistics. Imperial College Press, London, pp 25–44

One-Step Huber Estimates in the Linear Model

P. J. BICKEL*

Simple "one-step" versions of Huber's (M) estimates for the linear model are introduced. Some relevant Monte Carlo results obtained in the Princeton project [1] are singled out and discussed. The large sample behavior of these procedures is examined under very mild regularity conditions.

1. INTRODUCTION

In 1964 Huber [7] introduced a class of estimates (referred to as (M)) in the location problem, studied their asymptotic behavior and identified robust members of the group. These procedures are the solutions $\hat{\theta}$ of equations of the form,

$$\sum_{i=1}^{n} \psi(X_i - \hat{\theta}) = 0 \ , \tag{1.1}$$

where $X_1 = \theta + E_1, \cdots, X_n = \theta + E_n$ and $E_1, \cdots, E_n$ are unknown independent, identically distributed errors which have a distribution F which is symmetric about 0. If F has a density f which is smooth and if f is known, then maximum likelihood estimates if they exist satisfy (1.1) with $\psi = -f'/f$.

Under successively milder regularity conditions on ψ and F, Huber showed in [7] and [8] that such $\hat{\theta}$ were consistent and asymptotically normal with mean θ and variance $K(\psi, F)/n$ where

$$K(\psi, F) = \int_{-\infty}^{\infty} \psi^2(t) f(t) dt \Big/ \left[\int_{-\infty}^{\infty} f(t) d\psi(t) \right]^2 . \tag{1.2}$$

If F is unknown but close to a normal distribution with mean 0 and known variance in a suitable sense, Huber in [7] further showed that (M) estimates based on

$$\begin{aligned} \psi_K(t) &= t && \text{if} \quad |t| < K \\ &= K \operatorname{sgn} t && \text{if} \quad |t| \geq K \end{aligned} \tag{1.3}$$

have a desirable minimax robustness property. If K is finite these estimates can only be calculated iteratively. It has, however, been observed by Fisher, Neyman and others that if F is known and $\psi = (-f'/f)$, the estimate obtained by starting with a $\sqrt{n}$ consistent estimate $\tilde{\theta}$ and performing one Gauss-Newton iteration of (1.1) is asymptotically efficient even when the MLE is not and is equivalent to it when it is (cf. [13]). One purpose of this note is to show that under mild conditions this equivalence holds in the more general context of the linear model for general ψ.

Typically the estimates obtained from (1.1) are not scale equivariant.[1] To obtain acceptable procedures a scale equivariant and location invariant estimate of scale $\hat{\sigma}$ must be calculated from the data and $\hat{\theta}$ be obtained as the solution of

$$\sum_{j=1}^{n} \psi_{\hat{\sigma}}(X_j - \hat{\theta}) = 0 \ , \tag{1.4}$$

where

$$\psi_\sigma(x) = \psi(x/\sigma) \ . \tag{1.5}$$

The resulting $\hat{\theta}$ is then both location and scale equivariant. The estimate $\hat{\sigma}$ can be obtained simultaneously with $\hat{\theta}$ by solving a system of equations such as those of Huber's Proposal 2 [8, p. 96] or the "likelihood equations"

$$\begin{aligned} \sum_{j=1}^{n} \psi\left(\frac{X_j - \theta}{\sigma}\right) &= 0 \ , \\ \sum_{j=1}^{n} \chi\left(\frac{X_j - \theta}{\sigma}\right) &= 0 \ , \end{aligned} \tag{1.6}$$

where $\chi(t) = t\psi(t) - 1$. Or, we may choose $\hat{\sigma}$ independently. For instance, in this article, the normalized interquartile range,

$$\hat{\sigma}_1 = (X_{(n-[n/4]+1)} - X_{([n/4])})/2\Phi^{-1}(3/4) \ , \tag{1.7}$$

and the symmetrized interquartile range,

$$\hat{\sigma}_2 = \text{median}\ \{|X_j - m|\}/\Phi^{-1}(\tfrac{3}{4}), \tag{1.8}$$

are used where $X_{(1)} < \cdots < X_{(n)}$ are the order statistics, Φ is the standard normal cdf and m is the sample median. If $\hat{\sigma} \to \sigma(F)$ at rate $1/\sqrt{n}$ and F is symmetric as hypothesized, then the asymptotic theory for the location model continues to be valid with $K(\psi, F)$ replaced by $K(\psi(\frac{\cdot}{\sigma(F)}), F)$. (E.g., cf. [7].) We shall show (in the context of the linear model) under mild conditions that the one-step "Gauss-Newton" approximation to (1.4)—θ being the only unknown—behaves asymptotically like the root.

The estimates corresponding to ψ_K have a rather appealing form and, of course, all of these Gauss-Newton

* P.J. Bickel is professor, Department of Statistics, University of California, Berkeley, Ca. 94720. This research was performed with partial support of the O.N.R. under Contract N00014-67-A-D151-0017 with Princeton University, and N00014-67-A0114-0004 with the University of California at Berkeley, as well as that of the John Simon Guggenheim Foundation. The author would like to thank P.J. Huber, C. Kraft and C. Van Eeden and D. Relles for providing him with reprints of their work on this subject; W. Rogers III for programming the Monte Carlo computations of Section 3, which appeared in the Princeton project; and a referee who made Tables 1 and 2 reflect numerical realities.

[1] In this article location (scale) invariance refers to procedures which remain unchanged when the data are shifted (rescaled). The term "equivariant" is in accord with its usage in [2]. Thus, $\hat{\theta}$ location and scale equivariant means that $\hat{\theta}(aX_1 + b, \cdots, aX_n + b) = a\hat{\theta}(X_1, \cdots, aX_n) + b$ and $\hat{\sigma}$ scale equivariant means that $\hat{\sigma}(aX_1, \cdots, aX_n) = |a|\hat{\sigma}(X_1, \cdots, X_n)$.

June 1975, Volume 70, Number 350
Theory and Methods Section

procedures have the virtue of being simple and easily amenable to hand calculation for simple linear models. An analogous remark was made by Kraft and Van Eeden [11, 12] in connection with estimates based on rank tests.

Details of the model and the estimates are to be found in Section 2. Some Monte Carlo calculations are given in Section 3. Statements and proofs of the asymptotic behavior of the one-steps are given in Section 4. Finally, the proofs of some of the lemmas of Section 4 appear in an appendix.

2. THE MODEL AND ESTIMATES

The class of (M) estimates was extended to the general linear model by Relles [15] and Huber [9]. Here we observe $\mathbf{X} = (X_1, \cdots, X_n)$ where

$$X_j = \sum_{i=1}^{p} c_{ij}\beta_i + E_j , \quad 1 \le j \le n , \tag{2.1}$$

the E_j are as previously, the β_i unknown regression parameters and $C = \|c_{ij}\|$, the design matrix. An (M) *estimate* (scale equivariance not required) is defined quite naturally as a solution $\hat{\boldsymbol{\beta}} = (\hat\beta_1, \cdots, \hat\beta_p)$ of the system of equations

$$\sum_{j=1}^{n} c_{ij}\psi(Y_j(\boldsymbol{\beta})) = 0 , \quad 1 \le i \le p , \tag{2.2}$$

where

$$Y_j(\mathbf{t}) = X_j - \sum_{i=1}^{p} c_{ij}t_i \quad \text{if} \quad \mathbf{t} = (t_1, \cdots, t_p) . \tag{2.3}$$

Again, if $\psi = -f'/f$, these are the likelihood equations, and if $\psi(t) = t$, $\hat{\boldsymbol{\beta}} = \mathbf{X}C'[CC']^{-1}$, the least squares estimate. To obtain scale equivariance, we again need a scale equivariant estimate $\hat\sigma$ which is "shift" invariant, i.e.,

$$\hat\sigma(\mathbf{x} + \mathbf{t}C) = \hat\sigma(\mathbf{x}) . \tag{2.4}$$

The (scale equivariant) (M) estimates are now defined as the solutions of the system,

$$\sum_{j=1}^{n} c_{ij}\psi_{\hat\sigma}(Y_j(\hat{\boldsymbol{\beta}})) = 0 , \quad i = 1, \cdots, p . \tag{2.5}$$

Under various regularity conditions Relles and Huber [15, 9] have shown that $\hat{\boldsymbol{\beta}}$ is asymptotically normal with mean $\boldsymbol{\beta}$ and covariance matrix $K(\psi, F)[CC']^{-1}$ for the nonequivariant case and $K(\psi(\frac{\cdot}{\sigma(F)}), F)[CC']^{-1}$ otherwise. The efficiencies are independent of the design matrix and Huber's robustness results carry through. Let $\boldsymbol{\beta}^*$ be a given estimate of $\boldsymbol{\beta}$ which is shift equivariant, i.e.,

$$\boldsymbol{\beta}^*(\mathbf{x} + \mathbf{t}C) = \boldsymbol{\beta}^*(\mathbf{x}) + \mathbf{t} . \tag{2.6}$$

We shall say $\hat{\boldsymbol{\beta}}$ is a *one-step* (M) *estimate of Type 1* if ψ is absolutely continuous with derivative ψ' and $\hat{\boldsymbol{\beta}}$ satisfies the equations

$$\sum_{j=1}^{n} c_{ij}\psi(Y_j(\boldsymbol{\beta}^*)) = \sum_{k=1}^{p} (\hat\beta_k - \beta_k^*)\cdot\sum_{j=1}^{n} c_{kj}c_{ij}\psi'(Y_j(\boldsymbol{\beta}^*)), \quad 1 \le i \le p . \tag{2.7}$$

This system of equations is the linear approximation to the system (2.2) if we use $\boldsymbol{\beta}^*$ as an initial estimate. In the situations we are interested in, $\sum_{j=1}^{n} c_{kj}c_{ij}\psi'(Y_j(\boldsymbol{\beta}^*))$ is well approximated by its asymptotic expectation $\sum_{j=1}^{n} c_{kj}c_{ij}A(\psi, F)$, where

$$A(\psi, F) = \int_{-\infty}^{\infty} \psi'(t)dF(t) = -\int_{-\infty}^{\infty} f(t)d\psi(t) . \tag{2.8}$$

(The second equality holds only under mild regularity conditions.) The term on the right makes sense even when ψ is just of bounded variation on intervals. We shall use a slightly more general definition of $A(\psi, F)$ in (4.6). If $\hat A(\psi, F)$ is a consistent estimate of $A(\psi, F)$, we therefore define a *one-step* (M) *estimate of Type 2 as* the solution $\hat{\boldsymbol{\beta}}$ of the equations

$$\sum_{j=1}^{n} c_{ij}\psi(Y_j(\beta^*)) = \sum_{k=1}^{p} (\hat\beta_k - \beta_k^*)(\sum_{j=1}^{n} c_{kj}c_{ij})\hat A(\psi, F) , \tag{2.9}$$

or equivalently,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + \frac{1}{\hat A(\psi, F)} \cdot\{\psi(Y_1(\boldsymbol{\beta}^*)), \cdots, \psi(Y_n(\boldsymbol{\beta}^*))\}C'[CC']^{-1} \tag{2.10}$$

when CC' is nonsingular. Similarly we shall speak of scale equivariant one-step (ψ) estimates defined as previously, save that ψ is replaced by $\psi_{\hat\sigma}$ where $\hat\sigma$ is "shift" invariant *throughout.*

Our principal aim in introducing the one steps was to provide a version of Huber's estimate which is readily computable by hand in the location problem and other simple models. The ψ function of (2.2) here is given by (1.3). For a given scale estimate $\hat\sigma$ and the location model, the Type 1 one-step estimate may be written

$$\hat\beta = [\{\sum X_i : i \in S_0\} + K[N_+ - N_-]]/N_0 , \tag{2.11}$$

where $S_0 = \{i: |X_i - \beta^*| \le K\hat\sigma\}$, $S_+ = \{i: (X_i - \beta^*) > K\hat\sigma\}$, $S_- = \{i: (X_i - \beta^*) < -K\hat\sigma\}$ and N_0, N_+, N_- are the cardinalities of S_0, S_+ and S_-. If S_0 is empty the estimate is undefined. In the general case, let

$$S_0 = \{j: |X_j - \sum_{i=1}^{p} c_{ij}\beta_i^*| \le K\hat\sigma\} ,$$

etc. Then the Type 1 estimate is obtained as follows.

Replace any residual $X_j - \sum_{j=1}^{p} c_{ij}\beta_i^*$ by $K\hat\sigma$ if $j \in S^+$ and by $-K\hat\sigma$ if $j \in S^-$. If $j \notin S_0$, replace c_{ij} by 0 for $i = 1, \cdots, p$. If we denote the resulting vector of modified residuals by $\mathbf{R}^*$ and the resulting matrix of modified c_{ij} by C^*, then

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \mathbf{R}^*C'[C^*C^{*\prime}]^{-1} . \tag{2.12}$$

Alternatively, it is easy to see that if we define N_0 as before then under the conditions given in Section 4,

$$(N_0/n) \xrightarrow{P} A(\psi_\sigma, F) \tag{2.13}$$

and thus an alternative estimate (Type 2) would be

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + (n/N_0)\mathbf{R}^*C'[CC']^{-1} \tag{2.14}$$

Other possibilities are discussed in [9].

3. SOME MONTE CARLO RESULTS

As part of a larger study, [1], one-step estimates (for ψ_K) were considered as estimates for location under a variety of distributions and sample sizes. The following one-step procedures were considered. Let m denote the median, M the mean.

(1) $M15$; $K = 1.5$; $\hat{\sigma} = \hat{\sigma}_1$; $\tilde{\beta} = M$
(2) $D15$; $K = 1.5$; $\hat{\sigma} = \hat{\sigma}_1$; $\tilde{\beta} = m$
(3) $D20$; $K = 2.0$; $\hat{\sigma} = \hat{\sigma}_1$; $\tilde{\beta} = m$
(4) $P15$; $K = 1.5$; $\hat{\sigma} = \hat{\sigma}_2$; $\tilde{\beta} = m$

These were compared to the following Huber iterative estimates proposed by Hampel.

(5) $A15$; $K = 1.5$; $\hat{\sigma} = \hat{\sigma}_2$
(6) $A20$; $K = 2.0$; $\hat{\sigma} = \hat{\sigma}_2$

Note that comparison of $A15$ and $A20$ to $D15$ and $M15$ and $D20$, respectively, is reasonable since $\hat{\sigma}_2$ and $\hat{\sigma}_1$ are asymptotically equivalent to order $1/\sqrt{n}$ under mild regularity conditions, provided that F is symmetric.[2]

The sample sizes considered were $n = 5$, 10, 20 and 40. The distributions considered (not all being represented for each n) were:

(1) N—the normal
(2) C—the Cauchy
(3) 25 percent (NU)—a mixture of a standard normal distribution with the distribution of a standard normal variate divided by an independent variate having a uniform distribution on the interval (0, 1). The proportions were 75 percent normal, 25 percent of the latter distribution.
(4) t—the t distribution with three degrees of freedom.
(5) DE—the double exponential distribution.
(6) Pseudo-samples in which k observations were drawn from a normal distribution with variance nine (or 100) and the remaining $n - k$ were standard normal deviates. These are denoted by the notation

$$\left(\frac{k}{n}100\right) \text{ percent } \begin{cases} (3N) \\ (10N) \end{cases}.$$

Tables 1 and 2 were calculated using Exhibits 5.4–5.8 of [1] as well as measures of accuracy of these exhibits.[3] We refer to [1] for details of the Monte Carlo sampling procedure, a discussion of the accuracy of the results and other material of interest. Using between 640 and 1,000 replicates for each sample and some devices discussed in [1], essentially two-figure accuracy was obtained. We use the notation x/y to denote the efficiency of x with respect to y, i.e., the ratio $\mathrm{Var}\, y/\mathrm{Var}\, x$. Entries are 0 in cases such as those involving M or $M15$ under the C or 25 percent (NU) distribution in which the variances of these estimates are known to be infinite.

The asymptotic theory of Section 4 leads us to expect that $P15$ and $D15$ will behave like $A15$ and $D20$ like $A20$ in all of these cases. On the other hand, $M15$ should

1. Efficiencies of One Steps and Starting Points Versus Iterates for Sample Size 20

Efficiencies	*Distributions*						
	N	*25% (3N)*	*10% (10N)*	*DE*	t_3	*25% (NU)*	*C*
P15/A15	1.00	1.0	1.00	.99	1.0	1.0	1.0
m/A15	.70	.9	.83	1.13	.9	.9	1.5
D15/A15	1.00	1.0	.99	.96	1.0	1.0	.9
m/A20	.68	1.0	.92	1.22	1.0	1.0	1.9
D20/A20	1.00	1.0	.98	.97	1.0	1.0	.9
M/m	1.50	1.0	.16	.65	.6	0	0
M15/D15	1.00	1.0	.1−.3	1.01	.8	0	0

NOTE: For $n = 20$, the last significant figure is reliable at least up to ±1 for shapes other than 10 percent (10N) and up to ±2 for 10 percent (10N) unless a range is shown.

behave like $A15$ in Cases (1), (4), (5) and (6) only. What actually happened can be summarized as follows.

1. The difference between the one-step $P15$ and the iterate $A15$ set to the same scale is negligible across the whole range of distributions. However, the efficiency of the starting point m to $A15$ in this case is never less than .68.

2. If the starting point is too poor for the population at hand the loss in efficiency can be substantial. An example in point is t_3 where $M/m = .6$, $M15/D15 = .8$. Unfortunately, too few shapes and starting points were considered to see if there is a reasonable relation between the efficiency of the starting point to the iterate and that of the one step to the iterate.

3. The choice of scale has quite significant effects as the $P15/A15$, $D15/A15$ comparisons indicate. Unfortunately, the iterated forms of $D15$, $D20$ were not included in the study. Of course this has no bearing on the question of whether the one step is a good substitute for the iterated estimate.

4. Figures not included in this article but available in [1] indicate that the general qualitative nature of Tables 1 and 2 is unchanged if measures of spread other than the variance are used. However, the effect of a nonrobust starting point as in $M15$ is less severe.

5. The difference in computation time between iterate and one step can be substantial. In the Princeton study the average time of computation per estimate was recorded. From these figures it can be seen that the average percent increase in time for $A15$ versus $P15$ was of the order of 25 percent to 30 percent. (This is a percentage of the time required after all constants such as $\hat{\sigma}_2$ have been computed.) Preliminary computations for one steps with scale known for a standard Gaussian population ($n = 20$) indicate that the one-step starting at the median agrees with the iterate (up to two decimal places) between 80 ($K = 1.0$) and 60 ($K = 2.0$) percent of the time.

6. More extensive Monte Carlo computations need to be carried out to get a clear idea of the relationship between one-steps and iterates. This is particularly true for the smaller sample sizes for which the Princeton project figures are essentially unreliable.

[2] It is easy to show under symmetry that if $F' = f$ is finite at $F^{-1}(\frac{3}{4})$ then $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are asymptotically both Gaussian with mean $F^{-1}(\frac{3}{4})/\Phi^{-1}(\frac{3}{4})$ and variance $[4\Phi^{-1}(\frac{3}{4})f(F^{-1}(\frac{3}{4}))]^{-2}$. These assertions as well as asymptotic equivalence may be argued by replacing the quantile process by the empirical process as in Pyke-Shorack [14] or in the general linear model as in Bickel [4].

[3] Measures of accuracy of these exhibits do not appear in [1] but are available from Andrews *et al.*

2. Efficiencies of One Steps and Starting Points Versus Iterates for Sample Sizes 5, 10 and 40

Efficiencies	*n = 5*			*n = 10*				*n = 40*		
	N	*25% NU*	*C*	*N*	*20% 3N*	*25% NU*	*C*	*N*	*25%NU*	*C*
P15/A15	1.00	.8–1.2	.8–1.2	1.00	1.0	.9–1.0	.9–1.1	1.00	1.0	1.0
m/A15	.76	1.1–1.3	1.0–1.6	.77	1.0	.9–1.0	1.4–1.9	.68	.8	1.5
D15/A15	1.04	.6–.8	.5–1.1	1.02	.9	.2– .9	.1– .6	1.01	1.0	.9
m/A20	.73	1.0–1.5	1.1–1.9	.75	1.0	.9–1.0	1.8–2.3	.67	.8	1.9
D20/A20	1.02	.6–.9	.6–1.0	1.01	.9	.2– .9	.1– .6	1.06	1.0	.9
M/m	1.47	0	0	1.37	.7	0	0	1.53	0	0
M15/D15	.96	0	0	1.00	1.0	0	0	1.00	0	0

NOTE: For n = 5, 10 and shape N the last significant figure is reliable at least up to ±2. Otherwise unless a range is shown the last significant figure is reliable at least up to ±1.

4. THE LARGE SAMPLE BEHAVIOR OF ONE-STEPS

We shall prove asymptotic normality of the one-step estimates under the following simple conditions.

Condition G: The matrices CC'/n tend as $n \to \infty$ to a limit C_0 which is positive definite. Further,

$$\lim_n \max_{i,j} |c_{ij}|/\sqrt{n} = 0 \ . \tag{4.1}$$

We shall also need some smoothness conditions on ψ in addition to a consistency Condition A. The first set, labeled C, is appropriate for simple estimates while the second, S, is needed for scale equivariant estimates.

Condition A: $$\int \psi(t)dF(t) = 0 \ .$$

Clearly A holds if F is symmetric and ψ is antisymmetric.

Condition C1. The function ψ is of bounded variation in every interval, i.e., it may be written as

$$\psi = \psi^+ - \psi^- \tag{4.2}$$

where $\psi^\pm$ is monotone increasing and further,

$$\int_{-\infty}^{\infty} (\psi^\pm(x+h) - \psi^\pm(x-h))^2 dF(x) = O(1) \quad \text{as} \quad h \to 0 \ . \tag{4.3}$$

$$\sup \frac{1}{|h|} \left\{ \int_{-\infty}^{\infty} (\psi^\pm(x+q+h) - \psi^\pm(x+q))dF(x) : |q| \leqq \epsilon,\ |h| \leqq \epsilon \right\} < \infty \quad \text{for some } \epsilon > 0 \ . \tag{4.4}$$

Condition C2: Suppose that there exists $A(\psi^\pm, F)$ such that

$$\int_{-\infty}^{\infty} (\psi^\pm(x+h) - \psi^\pm(x))dF(x) = hA(\psi^\pm, F) + O(h) \ . \tag{4.5}$$

In this case define

$$A(\psi, F) = A(\psi^+, F) - A(\psi^-, F) \ . \tag{4.6}$$

Condition S_1: (a) The function ψ is as in (4.2) and

$$\sup \{(1/q^2) \int (\psi^\pm((1+\lambda+q)(x+h)) - \psi^\pm((1+\lambda)(x+h)))^2 dF(x) : |h| \leqq \epsilon,\ |\lambda| \leqq \epsilon,\ |q| \leqq \epsilon\} < \infty \tag{4.7}$$

for some $\epsilon > 0$.

$$\sup \left\{ 1/|h| \left| \int (\psi^\pm((1+\lambda)x+h) - \psi^\pm((1+\lambda)x-h))dF(x) \right| : |h| \leqq \epsilon,\ |\lambda| \leqq \epsilon \right\} < \infty \tag{4.8}$$

for some $\epsilon > 0$.

Condition S_2: There exists $A(\psi^\pm, F)$ such that

$$\int [\psi^\pm((1+\lambda)x+h) - \psi^\pm(x)]dF(x) = A(\psi^\pm, F)h + o(|h|) + O(|\lambda h|) + O(\lambda^2) \ . \tag{4.9}$$

Condition S_2 is satisfied if we can formally differentiate under the integral sign, ψ is antisymmetric and F is symmetric (about zero).

Finally we require further conditions on β^* and σ.

Condition B: If $\boldsymbol{\beta} = \mathbf{0}$,

$$\boldsymbol{\beta}^* = O_p(n^{-\frac{1}{2}}) \tag{4.10}$$

which by (2.6) implies that $\boldsymbol{\beta}^* - \boldsymbol{\beta} = O(n^{-\frac{1}{2}})$ in probability if $\boldsymbol{\beta}$ is true.

Condition D: There exists a positive functional $\sigma(F)$ such that

$$\hat{\sigma} = \sigma(F) + O_p(n^{-\frac{1}{2}}) \ . \tag{4.11}$$

(Because of the invariance assumption (2.4), Assertion (4.11) holds whatever be $\boldsymbol{\beta}$ if it holds for $\boldsymbol{\beta} = \mathbf{0}$.)

Moreover, writing σ for $\sigma(F)$, we shall suppose that

$$\hat{A}(\psi_{\hat{\sigma}}, F) \to A(\psi_\sigma, F) \tag{4.12}$$

in probability whatever be $\boldsymbol{\beta}$.

In the definitions and arguments which follow we shall assume that all probabilities and expectations are calculated under the assumption that $\boldsymbol{\beta} = \mathbf{0}$ unless the contrary is specifically indicated. Also let M be a generic constant. Define

$$T_n(\mathbf{t}) = \frac{1}{\sqrt{n}} \sum_{j=1}^n c_j[\psi(Y_j(\mathbf{t})) - E(\psi(Y_j(\mathbf{t}))] \ , \tag{4.13}$$

where we write c_j for c_{1j}.

Lemma 4.1: If G and C_1 hold, then

$$\sup \{|T_n(\mathbf{t}) - T_n(\mathbf{0})| : |\mathbf{t}| \leq M/\sqrt{n}\} \xrightarrow{p} 0 \ . \tag{4.14}$$

(We use $|\mathbf{t}|$ to denote the maximum of the absolute values of the coordinates of $\mathbf{t}$.) Let

$$T_n(\mathbf{t}, \lambda) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} c_j[\psi((1+\lambda)Y_j(\mathbf{t})) - E(\psi((1+\lambda)Y_j(\mathbf{t})))] \ . \quad (4.15)$$

Lemma 4.2: If G and S_1 hold, then

$$\sup \{ |T_n(\mathbf{t}, \lambda) - T_n(\mathbf{0}, 0)| : |\mathbf{t}| \leq M/\sqrt{n}, |\lambda| \leq \epsilon_n \} \xrightarrow{p} 0 \quad (4.16)$$

where $\epsilon_n \downarrow 0$ in any way whatever.

The proofs of these lemmas are given in the appendix.

From these lemmas we immediately get:

Proposition 4.1: (a) If G, C_1 and C_2 hold, then

$$\sup \left\{ \frac{1}{\sqrt{n}} |[\sum_{j=1}^{n} c_{1j}(\psi(Y_j(\mathbf{t})) - \psi(X_j)) + (\sum_{i=1}^{p} \sum_{j=1}^{n} c_{1j}c_{ij})A(\psi, F)]| : |\mathbf{t}| \leq \frac{M}{\sqrt{n}} \right\} \xrightarrow{p} 0 \ . \quad (4.17)$$

(b) If G, S_1 and S_2 hold, then

$$\sup \left\{ \frac{1}{\sqrt{n}} |[\sum_{j=1}^{n} c_{1j}(\psi((1+\lambda)Y_j(\mathbf{t})) - \psi(X_j)) + (\sum_{i=1}^{p} t_i \sum_{j=1}^{n} c_{1j}c_{ij})A(\psi, F)]| : |\mathbf{t}| \leq \frac{M}{\sqrt{n}}, |\lambda| \leq \frac{M}{\sqrt{n}} \right\} \xrightarrow{p} 0 \ . \quad (4.18)$$

Proof: Immediate upon expanding $E(\psi((1+\lambda)Y_j(\mathbf{t})) - \psi(X_j))$.

As an immediate consequence of this proposition we obtain

Theorem 4.1: If G, A, C_1, C_2, S_1, S_2, B and D hold and

$$\int \psi_\sigma^2(t)dF(t) < \infty$$

and $\hat{\boldsymbol{\beta}}$ is one step of Type 2, then under the model (2.1),

$$\sqrt{n}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - (\psi_\sigma(E_1), \cdots, \psi_\sigma(E_n))C'[CC']^{-1}[A(\psi_\sigma, F)]^{-1}\} \rightarrow 0 \quad (4.19)$$

in probability. A similar assertion holds when scale is not estimated. Hence, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ has a limiting Gaussian distribution with mean zero and covariance matrix $K(\psi_\sigma, F)C_0^{-1}$ where K is defined by (1.2) with the denominator in general given by $[A(\psi, F)]^2$.

Proof: By invariance reduce to the case $\boldsymbol{\beta} = \mathbf{0}$. Apply (4.18) with $\psi = \psi_\sigma$. Substitute $\beta_i^* - \beta_i$ for t_i, $(\hat{\sigma} - 1)$ for λ, c_{kj} for c_{1j}. Since $\sum_{i=1}^{p} (\beta_i^* - \beta_i) \sum_{j=1}^{n} c_{kj}c_{ij}$ is bounded in probability we can replace $A(\psi_\sigma, F)$ by $\hat{A}(\psi_{\hat{\sigma}}, F)$. The final result follows by Lindeberg's form of the central limit theorem.

Estimates of Type 1 satisfy the conclusion of Theorem 4.1 iff

$$\frac{1}{n\hat{\sigma}} (\psi_{\hat{\sigma}}'(Y_1(\boldsymbol{\beta}^*)), \cdots, \psi_{\hat{\sigma}}'(Y_n(\boldsymbol{\beta}^*)))CC' \xrightarrow{p} A(\psi_\sigma, F)C_0 \ . \quad (4.20)$$

It is easy to show that this is true if, in addition to our other conditions, either

Condition E_1: ψ' is uniformly continuous, or

Condition E_2: ψ' is of bounded variation in every interval and

$$E|[\psi']^{\pm}(aX_1 + b) - [\psi']^{\pm}(X_1)| = o(1) \quad \text{as} \quad a \rightarrow 1, b \rightarrow 0 \ .$$

Condition E_1 applies to smooth ψ while E_2 applies to Huber's ψ_K function. These conditions are far from necessary.

Although we have for completeness indicated the theory for the general linear model, in that context our theorem is best viewed as support for the feeling that a few iterations in solving a system of equations such as (2.5) lead to estimates whose behavior is much like that of the root. The reasons are:

1) For a multilinear regression, one usually employs a computer in any case and then solving the system (2.5) is not appreciably more difficult than obtaining the least squares estimates.
2) In such a case the only candidate for $\boldsymbol{\beta}^*$ is the least squares estimate, and as we shall see, even for moderately heavy tailed distributions the resulting one-step estimate can be poor.

However, for situations such as location, regression through the origin, and the c sample problem, where simple robust starting points such as the median or its analogues exist, the one-step estimates are easy to compute, and, as we have seen for location, quite satisfactory, at least if $\hat{\sigma}$ is chosen properly and the starting point is not too bad. Similar results hold if we replace Condition G by the more general

$$b^2(n)CC' \rightarrow C_0 \quad (4.21)$$

where $b(n) \rightarrow 0$, C_0 is positive definite,

$$b(n) \max_{i,j} |c_{ij}| \rightarrow 0 \ , \quad (4.22)$$

$\hat{\boldsymbol{\beta}}^*$ is $b^{-1}(n)$ consistent and all other conditions are unchanged.

If ψ is monotone rather than just of bounded variation it may be shown (see [16]) that these conditions guarantee convergence of the iterate as well as the one-step (M) estimate. If ψ is smooth and scale is known, it was shown by Huber [9] that a version of Theorem 4.1 holds for both iterates and one steps if $p \rightarrow \infty$ as well as n. The approach of this article does not extend readily to that case.

APPENDIX

Proof of Lemma 4.1: Without loss of generality take $\psi = \psi^+ \nearrow$. Begin by noting that for fixed $\mathbf{t}$ with $|\mathbf{t}| \leq M$,

$$T_n(\mathbf{t}/\sqrt{n}) - T_n(\mathbf{0}) \xrightarrow{p} 0 \ . \quad (A.1)$$

To see this, calculate

$$E\left(T_n\left(\frac{\mathbf{t}}{\sqrt{n}}\right) - T_n(0)\right)^2 = \frac{1}{n}\sum_{j=1}^n c_j^2 \operatorname{Var}\left(\psi\left(Y_j\left(\frac{\mathbf{t}}{\sqrt{n}}\right)\right)\right.$$
$$\left. - \psi(X_j)\right) \le \frac{1}{n}\sum_{j=1}^n c_j^2 \int_{-\infty}^{\infty}\left(\psi\left(s - \sum_{i=1}^p c_{ij}\frac{t_i}{\sqrt{n}}\right)\right.$$
$$\left. - \psi(s)\right)^2 f(s)ds \le \left\{\frac{1}{n}\sum_{j=1}^n c_j^2\right\} \max\left\{\int_{-\infty}^{\infty}(\psi(s+h)\right.$$
$$\left. - \psi(s))^2 f(s)ds \colon |h| \le pM \max_{i,j}|c_{ij}|/\sqrt{n}\right\} \to 0 \quad \text{(A.2)}$$

by Condition G and (4.3). Decompose the cube $K = \{\mathbf{t}\colon |\mathbf{t}| \le ([1/\delta] + 1)\delta M/\sqrt{n}\}$ as the union of cubes with vertices on the grid of points $(j_1\delta M/\sqrt{n}, \cdots, j_p\delta M/\sqrt{n})$ where the $j_i = 0, \pm 1, \cdots, \pm[1/\delta] + 1$. If $|\mathbf{t}| \le M/\sqrt{n}$, let $P(\mathbf{t})$ be (say) the lowest vertex of the cube containing $\mathbf{t}$. For fixed δ, by (A.2)

$$\max\{|T_n(P(\mathbf{t})) - T_n(0)|\colon |\mathbf{t}| \le M/\sqrt{n}\} \xrightarrow{p} 0 \ . \quad \text{(A.3)}$$

On the other hand, let K_1 be any cube of the partition and let P_1 be its lowest vertex. Then, by the monotonicity of ψ,

$$\sup\{|T_n(\mathbf{t}) - T_n(P_1)|\colon \mathbf{t} \in K_1\} \le \frac{1}{n}\sum_{j=1}^n |c_j|\left\{\left[\psi\left(Y_j(P_1) + \frac{M\delta}{\sqrt{n}}S_j\right) - \psi\left(Y_j(P_1) - \frac{M\delta}{\sqrt{n}}S_j\right)\right] + E\left[\psi\left(Y_j(P_1) + \frac{M\delta}{\sqrt{n}}S_j\right) - \psi\left(Y_j(P_1) - \frac{M\delta}{\sqrt{n}}S_j\right)\right]\right\} \quad \text{(A.4)}$$

where $S_j = \sum_{i=1}^p |c_{ij}|$. By arguing as for (A.2) it is easy to see that

$$\frac{1}{n}\operatorname{Var}\left\{\sum_{j=1}^n |c_j|\left[\psi\left(Y_j(P_1) + \frac{M\delta}{\sqrt{n}}S_j\right) - \psi\left(Y_j(P_1) - \frac{M\delta}{\sqrt{n}}S_j\right)\right]\right\} \to 0 \ . \quad \text{(A.5)}$$

It follows that to establish the lemma we need only check that

$$\max\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^n |c_j|\left[E\left(\psi\left(Y_j(P_n(\mathbf{t})) + \frac{M\delta}{\sqrt{n}}S_j\right)\right) - E\left(\psi\left(Y_j(P_n(\mathbf{t})) - \frac{M\delta}{\sqrt{n}}S_j\right)\right)\right]\colon |\mathbf{t}| \le \frac{M}{\sqrt{n}}\right\} = \delta O(1) \quad \text{(A.6)}$$

uniformly in δ, as $n \to \infty$.

Again using the monotonicity of ψ, it is clear that the expression in (A.6) is bounded by

$$\frac{1}{\sqrt{n}}\left[\sum_{j=1}^n |c_{ij}| \max\left\{E\left[\psi\left(X_1 + q + \frac{M\delta}{\sqrt{n}}S_j\right) - \psi\left(X_1 + q - \frac{M\delta}{\sqrt{n}}S_j\right)\right]\colon |q| \le \frac{MS_j}{\sqrt{n}}\right\}\right]$$

which by (4.4) is $O(\delta/n\sum_{i,j}|c_jc_{ij}|)$ uniformly in δ, for fixed M. The lemma follows.

Proof of Lemma 4.2: The estimate of (A.2) shows that if (4.16) holds,

$$T_n(\mathbf{t}_n, \lambda_n) = T_n(0, 0) + 0_p(1) \quad \text{(A.7)}$$

whenever $\mathbf{t}_n, \lambda_n \to 0$. Arguing as in Lemma 2.1 it is easy to see that it suffices to prove that

$$\sup\{|T_n(\mathbf{t}/\sqrt{n}, \lambda) - T_n(\mathbf{t}_0/\sqrt{n}, \lambda)|\colon |\mathbf{t} - \mathbf{t}_0| \le \delta,\ |\lambda| \le \epsilon_n\} = \delta 0_p(1) \quad \text{(A.8)}$$

uniformly in δ, and

$$\sup\{|T_n(\mathbf{t}_0/\sqrt{n}, \lambda) - T_n(\mathbf{t}_0/\sqrt{n}, 0)|\colon |\lambda| \leqq \epsilon_n\} = 0_p(1) \quad \text{(A.9)}$$

for each $\mathbf{t}_0$.

Now write

$$T_n(\mathbf{t}/\sqrt{n}, \lambda) = T_n(\mathbf{t}_0/\sqrt{n}, \lambda) + [T_n(\mathbf{t}/\sqrt{n}, \lambda) - T_n(\mathbf{t}_0/\sqrt{n}, \lambda)] \ . \quad \text{(A.10)}$$

Bound the last term, using the monotonicity of ψ as before, by

$$\frac{1}{\sqrt{n}}\sum_{j=1}^n |c_j|\left[\psi\left((1+\lambda)\left(X_j + \frac{\delta S_jM}{\sqrt{n}}\right)\right) - \psi\left(X_j + \frac{\delta S_jM}{\sqrt{n}}\right)\right]$$
$$+ \frac{1}{\sqrt{n}}\sum_{j=1}^n |c_j|\left[\psi\left(X_j - \frac{\delta S_jM}{\sqrt{n}}\right) - \psi\left((1+\lambda)\left(X_j - \frac{\delta S_jM}{\sqrt{n}}\right)\right)\right]$$
$$+ \frac{1}{\sqrt{n}}\sum_{j=1}^n |c_j|\left[\psi\left(X_j + \frac{\delta S_jM}{\sqrt{n}}\right) - \psi\left(X_j - \frac{\delta S_jM}{\sqrt{n}}\right)\right]$$
$$+ \frac{1}{\sqrt{n}}\sum_{j=1}^n |c_j| E\left[\psi\left((1+\lambda)\left(X_j + \frac{\delta S_jM}{\sqrt{n}}\right)\right) - \psi\left((1+\lambda)\left(X_j - \frac{\delta S_jM}{\sqrt{n}}\right)\right)\right]. \quad \text{(A.11)}$$

Let $W_n^{(i)}(\lambda, \delta)$, $1 \le i \le 3$, be the stochastic processes obtained by centering the preceding first three sums at their expectation. By (4.17) and Condition G,

$$E(W_n^{(1)}(\lambda_1, \delta) - W_n^{(1)}(\lambda_2, \delta))^2 \le K_1(\lambda_1 - \lambda_2)^2 \quad \text{(A.12)}$$

where K_1 is independent of n, λ_i, and δ for all λ_i sufficiently small, n large. Hence, by [5, p. 95],

$$\max\{|W_n^{(1)}(\lambda, \delta)|\colon |\lambda| \leqq \epsilon_n\} \xrightarrow{p} 0 \ . \quad \text{(A.13)}$$

A similar argument works for $W_n^{(2)}$, while $W_n^{(3)}$ may be taken care of as in the proof of Lemma 2.1. Similarly,

$$\max\{|T_n(\mathbf{t}_0/\sqrt{n}, \lambda) - T_n(\mathbf{t}_0/\sqrt{n}, 0)|\colon |\lambda| \le \epsilon_n\} \xrightarrow{p} 0 \quad \text{(A.14)}$$

uniformly in $|\mathbf{t}_0| \le M$. In view of (A.13) and (A.14), to prove (A.8) we need only bound

$$\frac{1}{\sqrt{n}}\sum_{j=1}^n |c_j|\left[E\left[\psi\left((1+\lambda)\left(X_j + \frac{\delta S_jM}{\sqrt{n}}\right)\right) - E\left(\psi\left((1+\lambda)\left(X_j - \frac{\delta S_jM}{\sqrt{n}}\right)\right)\right)\right]\right]$$

by $|\delta|O(1)$ for $|\lambda| \le \epsilon_n$. But this can be done using (4.18) as (4.14) was used in Lemma 2.1. Finally, (A.9) follows by using the same "tightness" argument as we employed for (A.13).

[Received September 1971. Revised July 1974.]

REFERENCES

[1] Andrews, D.F., *et al.*, *Robust Estimates of Location: Survey and Advances*, Princeton, N.J.: Princeton University Press, 1972.

[2] Berk, R., "A Special Structure and Equivariant Estimation," *The Annals of Mathematical Statistics*, 38 (October 1967), 1436–45.

[3] Bickel, P.J. and Wichura, M., "Convergence Criteria for Multi-parameter Stochastic Processes," *The Annals of Mathematical Statistics*, 42 (October 1971), 1656–70.

[4] ———, "Analogues of Linear Combinations of Order Statistics in the General Linear Model," *Annals of Statistics*, 1 (July 1973), 597–616.

[5] Billingsley, P., *Convergence of Probability Measures*, New York: John Wiley and Sons, Inc., 1968.

[6] Hàjek, J. and Sidak, Z., *Theory of Rank Tests*, New York: Academic Press, 1967.

[7] Huber, P.J., "The Behaviour of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium*, 1 (1965), 221–33.

[8] ———, "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35 (March 1964), 73–101.

[9] ———, "Robust Regression," *Annals of Statistics*, 1 (December 1973), 799–821.

[10] Kraft, C. and Van Eeden, C., "Efficient Linearized Estimates Based on Ranks," *Proceedings of the First International Symposium on Nonparametric Statistics*, Cambridge: Cambridge University Press, 1969, 267–73.

[11] ———, "Asymptotic Efficiencies of Methods of Computing Efficient Estimates Based on Ranks," *Journal of the American Statistical Association*, 67 (March 1969), 199–202.

[12] ———, "Linearized Rank Estimates and Signed Rank Estimates for the General Linear Hypothesis," *The Annals of Matehmatical Statistics*, 43 (January 1972), 42–57.

[13] Le Cam, L., "On the Asymptotic Theory of Estimation and Testing Hypotheses," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley: University of California Press, 1956, 129–56.

[14] Pyke, R. and Shorack, G., "Weak Convergence of a Two-Sample Empirical Process and a New Approach to the Chernoff-Savage Theorems," *The Annals of Mathematical Statistics*, 39 (June 1968), 755–71.

[15] Relles, D., "Robust Regression by Modified Least Squares," thesis, Yale University, 1968.

[16] Yohai, V.J., "Robust Estimation in the Linear Model." *Annals of Statistics*, 2 (May 1974), 562–67.

The Annals of Statistics
1984, Vol. 12, No. 4, 1349–1368

ROBUST REGRESSION BASED ON INFINITESIMAL NEIGHBOURHOODS[1]

BY P. J. BICKEL

University of California, Berkeley

We study robust estimation in the general normal regression model with random carriers permitting small departures from the model. The framework is that of Bickel (1981). We obtain solutions of Huber (1982), Krasker-Hampel (1980) and Krasker-Welsch (1982) as special cases as well as some new procedures. Our calculations indicate that the optimality properties of these estimates are more limited than suggested by Krasker and Welsch.

1. Introduction. Our aim in this paper is to compare and contrast robust regression estimates proposed by Huber (1973, 1982), Hampel (1978), Krasker (1978) and Krasker and Welsch (1982) as well as to derive and motivate other estimates using infinitesimal neighbourhood models as in Rieder (1978), Bickel (1981) for instance. Some of the results are stated in the discussion to Huber (1982) while others were presented at the 1979 Regression Special Topics Meeting in Boulder.

We consider a "stochastic" regression model. We observe (x_i, y_i), $i = 1, \cdots, n$ independent with common distribution P where the x_i are $1 \times p$, y_i scalar. We think of these observations as being obtained by contamination or some other stochastic perturbation from ideal but unobservable (x_i^*, y_i^*) which follow an ordinary Gaussian regression,

$$y_i^* = x_i^*\theta^T + u_i^*, \quad i = 1, \cdots, n$$

where the u_i^* are independent $\mathscr{N}(0, \sigma^2)$. Our aim is to estimate θ using the (x_i, y_i). For this formulation to make sense we must either:

(a) Specify P so that θ is identifiable. For instance let

$$x_i = x_i^* \quad \text{and} \quad y_i = x_i\theta^T + u_i$$

where the u_i are independent of x_i with common distribution symmetric about 0. This is the usual generalization of the linear model discussed e. g. in Huber (1973). For less drastic alternatives see Sacks and Ylvisaker (1978). This has the disadvantage of implicitly assuming that contamination conforms to the linear structure of the original model.

(b) Suppose that P is so close to the distribution P_0 of (x_i^*, y_i^*) that biases necessarily imposed by the lack of identifiability of θ are of the same order of magnitude as the standard deviations of good estimates. That is we assume P is

Received December 1982; revised January 1984.

[1] Research supported in part by Office of Naval Research Contract N00014-80-C-0163.

AMS 1980 *subject classifications.* Primary 62J05; secondary 62F35.

Key words and phrases. Regression, robustness, infinitesimal neighbourhoods, Krasker-Welsch estimates.

in "an order $1/\sqrt{n}$ neighbourhood" about P_0. By suitably choosing the metric defining the neighbourhood we can make precise our ideas about what departures we want to guard against as well as gauge the best that we can do against such departures in terms of classical decision theoretic measures such as M.S.E. For a general discussion of this point of view see Bickel (1981), hereafter [B]. This is the approach we take in this paper.

We apply this point of view to several types of neighbourhoods below and derive the optimal solutions. For regression through the origin we recapture the by now classical estimate of Hampel as well as Huber's (1982) MIA:A solution. For the general regression model we derive various natural extensions of the MIA:A procedure as well as the Hampel-Krasker and Krasker-Welsch procedures. Finally, we derive some negative results suggesting that the (1982) Krasker-Welsch conjecture is false.

Specifically, let $u_i = y_i - x_i\theta^T$, $i = 1, \cdots, n$. Suppose $\sigma^2 = 1$. Write $F = (G, H(\cdot \mid \cdot))$, $F_0 = (G_0, \Phi)$ where G, respectively G_0, is the marginal distribution of x_1, $H(\cdot \mid x)$ is the conditional distribution of u_1 given $x_1 = x$ and Φ is the standard normal distribution (of u_1^*). Since P and F determine each other we can describe neighbourhoods through conditions on $F, H(\cdot \mid \cdot)$. Such neighbourhoods, which will depend on n, will be denoted by $\mathscr{F}(t)$ (with subscripts) where $tn^{-1/2}$ is the size of the neighbourhood, $t \geq 0$.

Error-free x neighbourhoods: $G = G_0$ (or $x = x^*$).

Contamination: We suppose we can represent

$$H(\cdot \mid x) = (1 - \varepsilon(x))\Phi(\cdot) + \varepsilon(x)M(\cdot \mid x)$$

where $M(\cdot \mid x)$ is an arbitrary probability distribution. The contamination neighbourhoods $\mathscr{F}_{c0}(t)$, $\mathscr{F}_{ac0}(t)$ are completely specified by:

$$\mathscr{F}_{c0}(t)\colon \sup_x \varepsilon(x) \leq tn^{-1/2}, \quad \mathscr{F}_{ac0}(t)\colon \int \varepsilon(x)G_0(dx) \leq tn^{-1/2}.$$

That is, for both neighbourhoods the type of contamination of y for each x can be arbitrary. But under $\mathscr{F}_{c0}$ the conditional probability of contamination for each x is at most $tn^{-1/2}$ while under $\mathscr{F}_{ac0}$ only the marginal (or "average") probability of contamination is restricted. These are the types of departures considered by Huber (1982), Section 5.

Closely related are the metric neighbourhoods,

$$\mathscr{F}_{d0}(t)\colon \sup_x d(H(\cdot \mid x), \Phi) \leq tn^{-1/2}, \quad \mathscr{F}_{ad0}(t)\colon \int d(H(\cdot \mid x), \Phi)G_0(dx) \leq tn^{-1/2}$$

where d is a metric on the space of probability distributions on R. Of particular interest are the variational and Kolmogorov metrics given respectively by

$$v(P, Q) = \sup\{\, |P(A) - Q(A)| \colon A \text{ Borel}\},$$
$$k(P, Q) = \sup_x |P(-\infty, x] - Q(-\infty, x]|.$$

Recall that contamination neighbourhoods are contained in the corresponding variational neighbourhoods which are contained in the corresponding Kolmogorov neighbourhoods. The variational neighbourhoods can be interpreted as contamination neighbourhoods where ε can be a function not only of x but also of u^* and H is the conditional distribution of u_1 given x_1 and u_1^*. The complements of Kolmogorov neighbourhoods are identifiable in the sense of [B] at least if G_0 has finite support.

Errors in variables models: We drop the requirement that $G = G_0$ and proceed naturally, defining

$$\mathscr{F}_{c1}(t): F = (1 - \varepsilon)F_0 + \varepsilon M$$

where M is an arbitrary probability distribution on R^{p+1}, $\varepsilon = tn^{-1/2}$.

$$\mathscr{F}_{d1}(t): d(F, F_0) \leq tn^{-1/2}$$

where d is a metric on the probability distributions on R^{p+1}. Here v extends naturally and is of particular interest.

We consider estimates T_n of θ which are regression equivariant and asymptotically linear and consistent under the normal model. That is, for all $X_{n\times p}$, y, $b_{1\times p}$, T_n which is $1 \times p$ satisfies:

$$T_n(X, y + Xb^T) = T_n(X, y) + b \quad \text{(equivariance)} \tag{1.1}$$

and there exists ψ: $R^{p+1} \to R^p$ square integrable under F_0 such that

$$\int \psi(x, v)\Phi(dv)G_0(dx) = 0 \tag{1.2}$$

$$\int \psi^T(x, v)xv\Phi(dv)G_0(dx) = I, \quad \text{the } p \times p \text{ identity,} \tag{1.3}$$

and if $u = (u_1, \cdots, u_n)$, $X = (x_1^T, \cdots, x_n^T)^T$,

$$T_n(X, u) = n^{-1} \textstyle\sum_{i=1}^n \psi(x_i, u_i) + o_p(n^{-1/2}) \quad \text{(linearity and consistency)} \tag{1.4}$$

under F_0. Let $\Psi = \{\psi$: ψ square integrable function from R^{p+1} to R^p satisfying (1.2) and (1.3)$\}$.

All the usual consistent asymptotically normal estimates have this structure. In particular, under regularity conditions, the general (M) estimate T_n, solving

$$\textstyle\sum_{i=1}^n \psi(x_i, y_i - x_i T_n^T) = 0 \tag{1.5}$$

with $\psi \in \Psi$ satisfies (1.1) and (1.4). For members F of $\mathscr{F}$ leading to models contiguous to that given by F_0, (1.1)–(1.4) imply that $n^{1/2}(T_n - \theta)$ is asymptotically normal with mean

$$b(\psi, G, H) = n^{1/2} \int \psi(x, u)H(du \mid x)G(dx) \tag{1.6}$$

and variance-covariance matrix,

$$V(\psi) = \int \psi^T(x, u)\psi(x, u)\Phi(du)G_0(dx). \tag{1.7}$$

Note that b depends on n through G, H but for "regular" G, H stabilizes as $n \to \infty$.

In the univariate case, $p = 1$, we argue in [B] that we can characterize estimates which asymptotically minimize maximum (asymptotic) mean square error over $\mathscr{F}$ by minimizing $V(\psi) + \sup\{b^2(\psi, G, H): F \in \mathscr{F}\}$ over Ψ. More generally, the maximum risk of T_n as above, is for any reasonable symmetric loss function determined by $V(\psi)$ and $\sup\{|b(\psi, G, H)|: F \in \mathscr{F}\}$.

In Section 2 we study the univariate case as follows.

(1) We evaluate

$$b(\psi) = \lim \sup_n \sup\{|b(\psi, G, H)|: F \in \mathscr{F}\} \tag{1.8}$$

for the $\mathscr{F}$ we have introduced. Subscripts on b indicate which $\mathscr{F}$ we are considering.

(2) We solve the variational problem of minimizing $V(\psi)$ subject to $b(\psi) \leq m$. This is just Hampel's variational problem or a variation thereof.

The family of extremal $\{\psi_m : m \geq 0\}$ correspond formally via (1.5) to (M) estimates which are candidates for solutions to asymptotic min max problems. Checking that the (M) estimate or 1-step approximation to it actually is asymptotically minmax requires a uniformity argument such as that of Theorem 5, page 25 of [B] for the putative solution. These arguments are straightforward, requiring standard appeals to Huber (1967) or Bickel (1975) or Maronna and Yohai (1978). We therefore focus exclusively on the variational problems. No new procedures are obtained in this section. However, Theorem 2.1 formally gives some optimality properties of the Hampel and MIA:A estimates.

In Section 3 we consider the general multiple regression model and introduce WLS procedures and equivariance under change of basis in the independent variable space.

We derive various procedures on the basis of the optimality criteria we have advanced:

1) the Hampel-Krasker (nonequivariant) estimates;
2) the natural nonequivariant extension of Huber's MIA:A estimates (Theorem 3.1);
3) nonequivariant procedures which are also not WLS but are optimal for estimating one parameter at a time under $\mathscr{F}_{ac0}$;
4) an equivariant estimate which minimizes the maximum M.S.E. of prediction under $\mathscr{F}_{ac0}$ (Theorem 3.2);
5) the natural equivariant extension of Huber's MIA:A estimates which minimizes the maximum M.S.E. of prediction under $\mathscr{F}_{c0}$.

Finally we show that the optimality of the Hampel-Krasker and of the equivariant estimate minimizing the maximum M.S.E. of prediction depends on the quadratic form used in the loss function. This casts some doubt on a conjecture of Krasker and Welsch (1982). The doubt is confirmed by a recent counterexample of D. Ruppert.

2. Regression through the origin ($p = 1$). As we indicated, if $b(\psi)$ is given by (1.8), we want, for each $\mathscr{F}$ to solve the variational problem:

$$\text{(V)} \qquad \int \psi^2(x, u)\Phi(du)G_0(dx) = \min!$$

subject to (1.2), (1.3) and

$$b(\psi) \le m.$$

For each $\mathscr{F}$ we actually have a one-parameter family of variational problems as m varies and in principle each family could generate its own family of solutions. Fortunately there are only two families of solutions which we describe below.

It will be shown in Theorem 3.1 that for $\mathscr{F}$ which are of interest to us, only ψ which are Huber functions for each fixed x need be considered. That is, we can write ψ in the form:

$$(2.1) \qquad \begin{aligned} \psi(x, u) &= (a(x)/c(x))h(u, c(x)), & c(x) > 0 \\ &= a(x)\operatorname{sgn} u, & c(x) = 0 \end{aligned}$$

for given functions a; $c \ge 0$ satisfying (1.3) and $h(u, c) = \max(-c, \min(c, u))$.

For such ψ condition (1.2) is always satisfied and (1.3) becomes

$$(2.2) \qquad \int a(x)xB(c(x))G_0(dx) = 1$$

where

$$(2.3) \qquad B(c) = (2\Phi(c) - 1)/c \quad \text{with} \quad B(0) = 2\phi(0).$$

The two basic solution families of ψ which we denote $\{\psi_k\}$, $\{\tilde{\psi}_k\}$ will be defined by corresponding $\{a_k, c_k\}$, $\{\tilde{a}_k, \tilde{c}_k\}$ as follows:

For $0 < k < \infty$ let

$$(2.4) \qquad c_k(x) = k/|x|, \quad a_k(x) = \operatorname{sgn} x \Big/ \int (2\Phi(c_k(x)) - 1)x^2G_0(dx).$$

We add two limiting cases

$$(2.5) \qquad \psi_\infty(x, u) = xu \Big/ \int x^2G_0(dx)$$

$$(2.6) \qquad \psi_0(x, u) = \operatorname{sgn}(xu)/2\phi(0) \int |x| \, G_0(dx).$$

These are just the influence functions of the Hampel-Krasker-Welsch family of estimates. The extremal cases (2.5), (2.6) correspond to least squares, $T_n = \sum x_i y_i / \sum x_i^2$ and $T_n = \text{median } (y_i/x_i)$ respectively.

For $0 < t < 2\phi(0)$ let $0 < q(t) < \infty$ be the unique solution of

$$(2.7) \qquad 2(\phi(q) - q\Phi(-q)) = t.$$

Let $[2\underline{k}\phi(0)]^{-1}$ be the (G_0) ess sup of $|x|$. For $\underline{k} < k < \infty$ define

$$\tilde{c}_k(x) = q(1/k|x|)$$

$$\tilde{a}_k(x) = x \Big/ \int x^2(2\Phi(\tilde{c}_k(x)) - 1)I(|x| \geq [2k\phi(0)]^{-1})G_0(dx) \tag{2.8}$$
$$\text{if} \quad |x| \geq [2k\phi(0)]^{-1}$$
$$= 0 \text{ otherwise.}$$

The limiting cases are:

$$\tilde{\psi}_\infty(x, u) = \psi_\infty(x, u) \tag{2.9}$$

$$\tilde{\psi}_k(x, u) = \frac{k \operatorname{sgn} u}{\gamma}, \quad |x| = [2k\phi(0)]^{-1} \tag{2.10}$$
$$= 0 \qquad \text{otherwise}$$

if $\gamma = G_0\{x: |x| = [2k\phi(0)]^{-1}\} > 0$.

THEOREM 2.1. *Solutions to (V) are provided by*

(i) *Family* $\{\psi_k\}$: $\mathcal{F}_{\text{ac0}}$, $\mathcal{F}_{\text{av0}}$, $\mathcal{F}_{\text{ak0}}$, $\mathcal{F}_{\text{c1}}$, $\mathcal{F}_{\text{v1}}$, $\mathcal{F}_{\text{k1}}$

(ii) *Family* $\{\tilde{\psi}\}$ $\mathcal{F}_{\text{c0}}$, $\mathcal{F}_{\text{v0}}$, $\mathcal{F}_{\text{k0}}$

where we have substituted $d = v, k$ as appropriate in our notation. For given m, t the optimal k depends on m/t only and

(iii) *The solutions for* $\mathcal{F}_{\text{av0}}$, $\mathcal{F}_{\text{ak0}}$, $\mathcal{F}_{\text{v1}}$, $\mathcal{F}_{\text{k1}}$ *coincide.*

(iv) *The solutions for* $\mathcal{F}_{\text{v0}}$, $\mathcal{F}_{\text{k0}}$ *coincide.*

(v) *The solutions for* $\mathcal{F}_{\text{c0}}$ *are solutions for* $\mathcal{F}_{\text{v0}}$ *with m/t replaced by $m/2t$.*

The key to Theorem 2.1 is evaluation of $b(\psi)$ for the different neighbourhoods. The proof of a typical subset of the following assertions is given in the appendix.

If b is defined by (1.6), (1.8) then

$$b_{\text{c0}}(\psi) = t \int \text{ess sup}_u |\psi(x, u)| G_0(dx) \tag{2.11}$$

$$b_{\text{v0}}(\psi) = t \int [\text{ess sup}_u \psi(x, u) - \text{ess inf}_u \psi(x, u)] G_0(dx) \tag{2.12}$$

$$b_{\text{k0}}(\psi) = t \int \|\psi(x, \cdot)\| G_0(dx) \tag{2.13}$$

where "ess" refers to Lebesgue measure and $\|\cdot\|$ is the variational norm of $\psi(x, \cdot)$ viewed as a distribution function.

On the other hand,

$$b_{\text{c1}}(\psi) = t \text{ ess sup}_{x,u} |\psi(x, u)| \tag{2.14}$$

$$b_{\text{v1}}(\psi) = t[\text{ess sup}_{x,u}\psi(x, u) - \text{ess inf}_{x,u}\psi(x, u)] \tag{2.15}$$

(2.16) $$b_{\mathrm{k}}(\psi) = t \text{ ess sup}_x \| \psi(x, \cdot) \|.$$

The "average" models behave like "errors in variables".

(2.17) $$b_{a\cdot 0}(\psi) = b_{\cdot 1}(\psi).$$

If ψ is antisymmetric in u

(2.18) $$b_{\mathrm{v}i}(\psi) = 2b_{\mathrm{c}i}(\psi), \quad i = 0, 1.$$

If, in addition, ψ is monotone in u, then

(2.19) $$b_{\mathrm{k}i}(\psi) = b_{\mathrm{v}i}(\psi), \quad i = 0, 1.$$

Proof of Theorem. From (2.11)–(2.19) it is clear the solutions of (V) depend on m, t through m/t only and we can take $t = 1$. We claim it is enough to show (i) for $\mathscr{F}_{\mathrm{c}1}$, (ii) for $\mathscr{F}_{\mathrm{c}0}$. Since all members of both familes $\{\psi_s\}$ and $(\tilde{\psi}_s\}$ are antisymmetric and monotone in u, we can apply (2.18), (2.19) and the inclusion relations between the neighbourhoods to derive (iii)–(iv). From (iii)–(iv), (i) and (ii) follow for all neighbourhoods and (v) is immediate.

Problem (V) for $\mathscr{F}_{\mathrm{c}1}$ is just Hampel's variational problem. Existence of a solution follows from standard weak compactness arguments. For these and the derivation of the family of solutions by a standard Lagrange multiplier argument, see, for example, [B].

Problem (V) for $\mathscr{F}_{\mathrm{c}0}$ is a little less standard. Huber (1982) essentially derives the solution indirectly from his finite minimax robust testing theory.

We will give another proof which relies on a "conditional on x" Lagrange multiplier argument for the p-variate case. See the proof of Theorem 3.1 and note (2) following it. □

Discussion.

(1) *Unknown G_0.* In practice G_0 is unknown. Strictly speaking it is not required for the calculation of any particular estimate of the families $\{\psi_k\}$, $\{\tilde{\psi}_k\}$. However, in order to pick out a member on optimality grounds, say, minimizing maximum M.S.E., and to estimate maximum M.S.E., G_0 is required. Estimating G_0 by the empirical distribution of the x_i gives the same asymptotic results.

(2) *Unknown scale.* In practice the scale σ^2 of the u_i^* is unknown. As we indicate in [B] under mild conditions, the estimate T_n solving

(2.20) $$\textstyle\sum_{i=1}^n \psi(x_i, (y_i - x_i T_n)/s) = 0$$

where s is a consistent estimate of σ (over $\mathscr{F}$) and ψ is antisymmetric in u for fixed x will have influence function $\sigma\psi(x, u/\sigma)$. It follows that the optimal ψ functions derived under the assumption σ known can be modified as in (2.20) to yield estimates optimal whatever be σ. There are serious questions of computation and existence of solutions when scale is estimated simultaneously. See Maronna (1976) and Krasker and Welsch (1982).

(3) The agreement between the errors in variables and average c or v models

is interesting though, in retrospect, not surprising. As Huber (1982) reveals for the average c model, Nature can be thought of as using most of her allocated ε of contamination to create very skew conditional given x distributions of u for the largest x and this can certainly also be done for errors in variables.

(4) The qualitative behaviour for $\mathscr{F}_{c0}$ (and $\mathscr{F}_{v0}$) is surprising as noted by Huber (1982). Small x's which are relatively uninformative are cut out by the $\tilde{\psi}$ estimates and on the other hand the $\tilde{\psi}$ are not bounded. (However if G_0 is estimated as it must be by the empirical d.f. of the x_i, $\sup_{i,u} | \tilde{\psi}_k(x_i, u) | < \infty$ for each n.) In this case since Nature is required to spread her contamination evenly, it pays to take chances and use c large at the large values of x which are informative if they are not contaminated and it does not pay to take any chances at the small and uninformative values of x.

(5) Interestingly enough, the same behaviour is exhibited by the Hellinger metric neighbourhoods $\mathscr{F}_{h0}$ where $h^2(P, Q) = \int (\sqrt{dP/du} - \sqrt{dQ/du})^2\, du$. Here it may be shown

$$b_{h0}(\psi) = 2t \int \left(\int \psi^2(x, u)\Phi(du) \right)^{1/2} G_0(dx)$$

and the resulting optimal ψ are of the form

$$\psi_k^*(x, u) = a(x)u$$

where

$$\begin{aligned} a(x) &= 0, & |x| &\le k \\ &= \mu(x - k \operatorname{sgn} x), & |x| &> k, \end{aligned}$$

where μ is determined by (1.3).

These solutions do not agree with the unique solution $\psi_\infty(x, u)$ (essentially least squares), appropriate for $\mathscr{F}_{ah0}$, $\mathscr{F}_{h1}$.

3. The general case. For $p > 1$ we face the usual problem of choosing adequate scalar summaries (measures of loss) of the vector $b(\psi, F)$ and the matrix $V(\psi)$ on which to optimize.

Again ψ's which are Huber functions for each x play a special role,

$$\psi(x, u) = (a(x)/c(x))h(u, c(x)) \tag{3.1}$$

where a is now a vector, $c \ge 0$. For such ψ, (1.2) is satisfied, (1.3) becomes

$$\int x^T a(x) B(c(x)) G_0(dx) = I \tag{3.2}$$

and

$$V(\psi) = \int a^T a(x) A(c(x)) G_0(dx) \tag{3.3}$$

where

$$A(c) = \frac{2\Phi(c) - 1 - 2c\phi(c)}{c^2} + 2\Phi(-c), \quad A(0) = 1. \tag{3.4}$$

Also natural are ψ corresponding to weighted least squares estimates (WLS) definable in the multivariate case by

$$T_n = \textstyle\sum_{i=1}^n w_i y_i x_i (\sum_{i=1}^n w_i x_i^T x_i)^{-1}$$

with

$$w_i = w(x_i, y_i - x_i T_n^T)$$

scalars defined up to a proportionality constant. Note that ψ corresponds to a WLS estimate $\Leftrightarrow$ the direction of ψ is that of a linear transformation of x, i.e.,

$$\psi(x, u) = w(x, u)uxR \tag{3.5}$$

with

$$R^{-1} = \int x^T x w(x, u) u^2 \Phi(du) G_0(dx).$$

We classify solutions to the p-variate problem according as they do or do not possess equivariance under changes of basis in the X-space. An estimate T_n is equivariant under change of basis if and only if

$$T_n(XB, y) = T_n(X, y)[B^T]^{-1}.$$

(a) *Nonequivariant solutions.*

(i) *The Hampel-Krasker solution.* Perhaps the most natural choice of objective function is the total M.S.E. of the components, tr $V(\psi) + bb^T(\psi, F)$. If we let $|\cdot|$ denote the Euclidean norm, this leads to the following p-variate version of (V),

$$\int |\psi|^2(x, u)\Phi(du)G_0(dx) = \text{min!} \tag{V}$$

for $\psi \in \Psi$ and $\sup_{\mathscr{F}} |b|(\psi, F) \le m$. Holmes (1982) has shown that for $\mathscr{F}_{\text{ac}0}, \mathscr{F}_{\text{c}1}$,

$$\sup_{\mathscr{F}} |b|(\psi, F) = t \operatorname{ess\,sup}_{x,u} |\psi(x, u)|$$

so that (V) is just the problem of Krasker, Hampel (1978) whose solution is of the form, for $\lambda_0 < \lambda < \infty$,

$$\psi(x, u, \lambda) = xQh(u, \lambda/|xQ|)$$

where Q is symmetric positive definite and by (3.2)

$$Q^{-1} = \int x^T x\left(2\Phi\left(\frac{\lambda}{|xQ|}\right) - 1\right)G_0(dx).$$

Here

$$\lambda = \text{ess sup}_{x,u} | \psi(x, u, \lambda) |$$

and

$$0 < \lambda_0 = \inf\{\sup_{x,u} | \psi(x, u) |: \psi \in \Psi\}.$$

The solution to (V) has $\lambda = mt$. Krasker and Hampel (see also [B]) show that whenever there exists ψ with $\text{ess sup}_{x,u} | \psi(x, u) | = \lambda > \lambda_0$, then $\psi(\cdot, \cdot, \lambda)$ exists and is unique.

Note that $\psi(\cdot, \cdot, \lambda)$ is of the form (3.1) and also WLS with

$$a(x) = \lambda(xQ/|xQ|), \quad c(x) = \lambda/|xQ|, \quad w(x, u) \propto h(u, c(x))/u.$$

NOTES.

(1) Calculations along the lines of Maronna (1976) show that $\lambda \to Q_\lambda$ is decreasing (in the order on positive definite symmetric matrices).

(2) It may be shown that $\lambda_0 \geq p/2\phi(0) \int |x| G_0(dx)$.

(ii) *A generalization of Huber's approach.* For $\mathscr{F}_{c0}$ it seems difficult to evaluate $\sup_{\mathscr{F}} |b|(\psi, F)$ exactly. However, it is easy to show that (see appendix)

$$\sup\{|b|(\psi, F): F \in \mathscr{F}_{c0}\} \leq t \int \sup_u | \psi(x, u) | G_0(dx).$$

As in the 1-dimensional case $\int \sup_u | \psi(x, u) | G_0(dx)$ can be interpreted as an average sensitivity. The solution of the resulting problem,

$$\text{(V}'\text{)} \qquad \int | \psi(x, u) |^2 \Phi(du) G_0(dx) = \min!$$

subject to (1.2), (1.3) and

$$\int \sup_u | \psi(x, u) | G_0(dx) \leq \lambda$$

for $\lambda = m/t$, yields what should be a reasonable approximation to (V).

THEOREM 3.1. *For every $\lambda > \lambda_1$ there exists a unique pair $(s(\lambda), Q(\lambda))$ such that*

$$\tilde{\psi}(\cdot, \lambda) = \rho(\cdot, Q(\lambda), s(\lambda))$$

is an influence function and

$$\text{(3.6)} \qquad \int \sup_u | \tilde{\psi}(x, u, \lambda) | G_0(dx) = \lambda$$

and $\tilde{\psi}(\cdot, \lambda)$ solves (V′).

The solutions to (V′) are describable as follows: Define, for $s > 0$, Q symmetric positive definite, q as in (2.7),

$$\begin{aligned}\rho(x, Q, s) &= xQh(u, q([s\,|xQ|]^{-1})), \quad |xQ| > [2s\phi(0)]^{-1}\\ &= 0 \qquad\qquad\qquad\qquad \text{otherwise.}\end{aligned}$$

Let

$$\lambda_1 = \inf\left\{\int \sup_u |\psi(x, u)|\, G_0(dx) \colon \psi \in \Psi\right\}.$$

$\tilde{\psi}(\cdot, \lambda)$ can be written in the form (3.1) with corresponding functions defined for $s = s(\lambda)$, $Q = Q(\lambda)$ by

$$\begin{aligned}\tilde{c}(x, \lambda) &= q(|sxQ|^{-1})\\ \tilde{a}(x, \lambda) &= xQ\tilde{c}(x, \lambda) \qquad \text{for} \quad |xQ| > [2s\phi(0)]^{-1}\\ &= 0 \qquad\qquad\qquad \text{otherwise.}\end{aligned}$$

Preliminary calculations along the lines of Maronna (1976) and Maronna-Yohai (1981) indicate that at least if G_0 does not place mass on hyperplanes, then Q is uniquely determined by s through (3.2), i.e.

$$Q^{-1} = \int_{S(s,Q)} x^T x(2\Phi(q(|sxQ|^{-1})) - 1)G_0(dx) \tag{3.7}$$

where $S(s, Q) = \{x \colon |sxQ| > 2\phi(0)\}$ and then s is determined by λ through (3.6)

$$\int_{S(s,Q)} |xQ|\, q(|sxQ|^{-1})G_0(dx) = \lambda. \tag{3.8}$$

Moreover if we write Q_s for the solution of (3.7), $s \to Q_s$ is nondecreasing and hence $\lambda \to s(\lambda)$ is also. So we can reparametrize $\tilde{\psi}(\cdot, \lambda)$ by s for $s > \inf\{s(\lambda) : \lambda > \lambda_1\}$. If, for $p = 1$, we take $k = sQ_s$, then we obtain the family $\tilde{\psi}_k$ of Theorem 2.1. Since k is an increasing function of λ we obtain the conclusions of Theorem 2.1.

PROOF. In the appendix we show by standard optimization theory arguments that a solution to (V′) exists and is also the solution to a Lagrangian problem

$$\int \left\{|\psi|^2(x, u) - 2\int u\psi(x, u)Qx^T + \frac{2}{s}|\psi|(x, u)\right\}\Phi(du)G_0(dx) = \min!$$

for $Q_{p\times p}$, $s > 0$.

If ψ_0 is the solution we can minimize

$$\int |\psi|^2(x, u)\Phi(du) - 2\int u\psi(x, u)Qx^T\Phi(du)$$

subject to $\sup_u |\psi(x, u)| \le \sup_u \psi_0(x, u)$ and conclude that ψ_0 is of the form (3.1) with the corresponding vector $a_0(x)$ and $c_0(x)$ minimizing

$$\int \{|a|^2(x)A(c(x)) - 2xQa^T(x)B(c(x)) + s^{-1}|a|(x)\}G_0(dx).$$

Minimizing pointwise we obtain as necessary conditions for a_0, c_0

$$a_0A(c_0) = xQB(c_0) + s^{-1}(a_0/|a_0|) = 0, \quad a_0 \neq 0 \tag{3.9}$$

$$\begin{aligned} |a_0|^2 &\leq xQa_0^Tc_0 \\ &= xQa_0^Tc_0 \quad \text{if} \quad c_0 > 0. \end{aligned} \tag{3.10}$$

From (3.10), $a_0 \neq 0 \Rightarrow c_0 > 0$. Then by (3.9)

$$a_0 = |a_0|(xQ/|xQ|) = c_0xQ$$

by (3.10). Again by (3.9)

$$c_0A(c_0) - B(c_0) + (1/s|xQ|) = 0$$

which implies $|xQ| \geq [2s\phi(0)]^{-1}$, $c_0 = q([s|xQ|]^{-1})$. Conversely, if $|x| > [2s\phi(0)]^{-1}$, $\tilde{a}(x, \lambda)$, $\tilde{c}(x, \lambda)$ yield

$$|a|^2A - 2xQa^TB(c) + s^{-1}|a| < 0$$

and hence $0 \neq a_0 = \tilde{a}$ by our previous reasoning. Since $\tilde{\psi}$ must satisfy (1.2), Q must satisfy (3.9) and be positive definite symmetric. The theorem is proved. □

(iii) *One at a time optimality.* Another nonequivariant solution of interest is obtained by minimizing the maximum M.S.E. of each component of θ separately. That is, we seek $\psi^* = (\psi_1^*, \cdots, \psi_p^*) \in \Psi$ which *simultaneously* minimizes

$$\int [\psi_j]^2(x, u)\Phi(du)G_0(dx)$$

for $\psi = (\psi_1, \cdots, \psi_p) \in \Psi$ and

$$\sup\{|b_j(\psi, F)|: F \in \mathscr{F}\} \leq m_j$$

where $b(\psi, F) = (b_1(\psi, F), \cdots, b_p(\psi))$. For neighbourhoods of the "average" or errors in variables types, the solutions ψ^*, indexed by the vector $m = (m_1, \cdots, m_p)$, are *not* of the WLS form. They are given by

$$\psi_j^*(x, u; m) = uxa_j^Th(u, m_j/|xa_j^T|), \quad j = 1, \ldots, p \tag{3.11}$$

where (1.2) and (1.3) hold. Existence of $\psi^*(\cdot, m_0)$ and their form as solutions of a Lagrange problem are guaranteed. for m_0 an interior point of $\{m: t \sup_{x,u}|\psi_j(x, u)| \leq m_j, j = 1, \cdots, p\}$. The limiting case corresponding to the median is, for $x = (x_1, \cdots, x_p)$,

$$\psi_j^*(x, u) = c_j\operatorname{sgn}[(x_j - \textstyle\sum_{k\neq j} b_{kj}x_k)u] \tag{3.12}$$

where

$$c_j = \left[\left(\frac{2}{\pi}\right)^{1/2} \int |x_j - \textstyle\sum_{k\neq j} b_{kj}x_k| G_0(dx)\right]^{-1}$$

where $B = \| b_{ij} \|$ is determined by

$$\int \operatorname{sgn}(x_j - \textstyle\sum_{k \neq j} b_{kj} x_k) x_i G_0(dx) = 0, \quad i \neq j. \tag{3.13}$$

If $(x_{i1}, \cdots, x_{ip}, y_i)$, $i = 1, \cdots, p$ are the observations, $\hat{\theta}_1, \cdots, \hat{\theta}_p$ are the estimates, and $\hat{\varepsilon}_i = y_i - \sum_{j=1}^p x_{ij}\hat{\theta}_j$ are the residuals, then $\hat{\theta}_1, \cdots, \hat{\theta}_p$ are characterized by the property that

$$\operatorname{median}_i \hat{\varepsilon}_i / (x_{ij} - \textstyle\sum_{k \neq j} b_{kj} x_{ik}) \cong 0$$

for $j = 1, \cdots, p$. In view of (3.13) the b_{kj} can be interpreted as the coefficients of a least absolute residuals fit of $\sum_{k \neq j} b_k x_k$ to x_j, i.e.,

$$\int | x_j - \textstyle\sum_{k \neq j} b_{kj} x_k | G_0(dx) = \min \int | x_j - \textstyle\sum_{k \neq j} b_k x_k | G_0(dx). \tag{3.14}$$

This characterization guarantees the existence of this influence function at least if G_0 is absolutely continuous. Of course, there may be difficulties for a sample where we replace G_0 by the empirical d.f. of the X_i.

At first glance this solution appears to render the Hampel-Krasker solution inadmissible. This is, however, not the case. ψ^* here minimizes (for suitable m_j),

$$R(\psi) = \textstyle\sum_{i=1}^p \displaystyle\int \psi_i^2(x, u)\Phi(du)G_0(dx) + \textstyle\sum_{i=1}^p \max_{\mathscr{F}} b_i^2(\psi, F)$$

while the Hampel-Krasker solution minimizes

$$S(\psi) = \textstyle\sum_{i=1}^p \displaystyle\int \psi_i^2(x, u)\Phi(du)G_0(dx) + \max_{\mathscr{F}} \textstyle\sum_{i=1}^p b_i^2(\psi, F).$$

Of course, $S \leq R$ but the optimal solutions are not related.

(b) *Equivariant solutions.* When translated to influence functions this equivariance becomes

$$\psi(x, u, G_0) = \psi(xB, u, G_0 B^{-1})B^T \tag{3.15}$$

where $\psi(x, u, G)$ is the influence curve if $X_1 \sim G$.

(i) *Equivariant best MSE of prediction.* Suppose that X_1 is error free so that $G = G_0$ and that $\int | x |^2 G_0(dx) < \infty$. The most natural way of obtaining invariant ψ with local optimality properties is to use as objective function the expected mean square error of prediction

$$\int \{xV(\psi)x^T G(dx) + xb^T(\psi)b(\psi)x^T\} G_0(dx).$$

We can rewrite this as

$$\int \psi\Sigma\psi^T(x, u)\Phi(du)G_0(dx) + b(\psi, F)\Sigma b^T(\psi, F)$$

where

$$\Sigma = \int x^T x G_0(dx). \tag{3.16}$$

As in the noninvariant case we can deal easily with $\mathscr{F}_{a0c}$ since

$$\sup\{b(\psi, F)\Sigma b^T(\psi, F): F \in \mathscr{F}_{ac0}\} = \text{ess sup}_{x,u}\psi(x, u)\Sigma\psi^T(x, u). \tag{3.17}$$

Minimizing the maximum of our objective function over $\mathscr{F}_{a0c}$ is easy once we have solved

$$\int \psi\Sigma\psi^T(x, u)\Phi(du)G_0(dx) = \min! \tag{V$_\mathrm{I}$}$$

for $\psi \in \Psi$ such that

$$\text{ess sup}_{x,u}\psi\Sigma\psi^T(x, u) \le \lambda.$$

Let

$$\lambda_{I0} = \inf \text{ess}\{\sup_{x,u}\psi\Sigma\psi^T(x, u): \psi \in \Psi\}$$

$$d^2(x, \Sigma) = x\Sigma x^T.$$

For $\lambda > \lambda_{I0}$ let

$$\psi_I(x, u, \lambda) = xQh(u, \lambda/d(xQ, \Sigma)) \tag{3.18}$$

where Q is positive definite symmetric,

$$\int x^T x\left(2\Phi\left(\frac{\lambda}{d(xQ, \Sigma)}\right) - 1\right)G_0(dx) = Q^{-1}. \tag{3.19}$$

THEOREM 3.2. *If* $\lambda > \lambda_{I0}$, $\psi_I(\cdot, \cdot, \lambda)$ *uniquely solves* (V$_\mathrm{I}$).

PROOF. Again by standard arguments we can establish existence of a minimizing ψ_0 which solves an equivalent Lagrangian problem

$$\int \{\psi\Sigma\psi^T(x, u) - 2\int uxQ\Sigma\psi^T(x, u)\}\Phi(du)G_0(dx) = \min!$$

subject to $|\Psi\Sigma\psi^T| \le \lambda$. A direct minimization of $\psi\Sigma\psi^T - 2uxQ\Sigma\psi^T$ under the side condition yields (3.18) and (3.2) implies (3.19). □

Note that the uniqueness of ψ_I and (3.19) imply the equivariance property (3.15).

(ii) *An equivariant Huber solution.* As in the nonequivariant case we can bound the maximum expected squared bias of the predictor

$$\sup\left\{\int xb^T b(\psi, F)x^T G_0(dx): F \in \mathscr{F}_{c0}\right\}$$

above by

$$t \int \{\sup_u \psi(x, u) \Sigma \psi^T(x, u)\} G_0(dx).$$

The resulting variational problem

$$\int \psi \Sigma \psi^T(x, u) \Phi(du) G_0(dx) = \min!$$

subject to

$$\int \sup_u \psi(x, u) \Sigma \psi^T(x, u) G_0(dx) \le \lambda \tag{3.20}$$

has solutions of the form

$$\tilde{\psi}(x, u, s) = \frac{\tilde{a}_I(x, s)}{\tilde{c}_I(x, s)} h(u, \tilde{c}_I(x, s)) \tag{3.21}$$

where

$$\tilde{c}_I(x, \lambda) = q(1/sd(xQ, \Sigma)), \quad \tilde{a}_I(x, s) = xQ\tilde{c}_I(x, s)$$

if

$$\begin{aligned} d(xQ, \Sigma) &\ge [2s\phi(0)]^{-1} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

and Q, s are determined by the requirement that $\tilde{\psi}_I$ is an influence function satisfying equality in (3.20).

Reparametrizations are possible for the procedures of this section as for the Hampel-Krasker and Huber solutions.

(iii) *The Krasker-Welsch* (1982) *solution.* Based on sensitivity considerations, Krasker and Welsch proposed estimates given by

$$\psi_{KW}(x, u, \lambda) = xQh(x, \lambda/d(xQ, V^{-1})), \quad \lambda > \sqrt{p} \tag{3.22}$$

where

$$\int x^T x \left(2\Phi\left(\frac{\lambda}{d(xQ, V^{-1})} \right) - 1 \right) G_0(dx) = Q^{-1}$$

and

$$V_\lambda = \int \psi^T \psi(x, u, \lambda) \Phi(du) G_0(dx). \tag{3.23}$$

Equivalently if $A^{-1} = QV^{-1}Q$, (3.23) becomes

$$A = \int x^T x [2\Phi(\lambda/d(x, A^{-1})) - 1 - 2\lambda d^{-1}(x, A^{-1}) \phi(\lambda d^{-1}(x, A^{-1}))] G_0(dx)$$

and Q may be obtained directly from (3.22). Existence of the K-W solution for

$\lambda > \sqrt{p}$ is guaranteed by results of Maronna (1976). The K-W solution is also equivariant. It evidently has the property (by arguing as for Theorem 3.2) of uniquely minimizing $\int \psi V^{-1}(\psi_{KW})\psi^T$ subject to $\sup \psi V^{-1}(\psi_{KW})\psi^T \le \lambda^2$. Krasker and Welsch conjecture a strong optimality property (see below).

(iv) *More general optimality properties.* Whatever be p, least squares estimates do not minimize only trace $V(\psi)$ but the matrix itself or equivalently $\int \psi M \psi^T$ for all M positive definite, symmetric. It is fairly easy to see (see also Stahel, 1981) that once we bound the vector influence curve as we have in this section, no such conclusion is possible. Thus $\psi M \psi^T(x, u) - 2u\psi(x, u)QMx^T$ is minimized subject to $|\psi| \le \lambda$ by $\psi = uxQ$ if $|u| \le \lambda/|xQ|$, but, unless $M = I$, by a boundary value other than $\lambda(xQ/|xQ|)$ if $|u| > \lambda/|xQ|$.

Krasker and Welsch seek to remedy this failing by restricting ψ to the WLS form, i.e., forcing the direction of ψ to coincide with a linear transformation of x. They conjecture that their solution minimizes $V(\psi)$ among all WLS estimates with $\sup \psi V^{-1}(\psi)\psi^T \le \eta$. Our methods do not readily give a counterexample to their conjecture but we show below that neither the Hampel-Krasker estimate nor the equivariant estimate of section (i) possess the analogous optimality property, thus casting some doubt on the conjecture. (David Ruppert has recently discovered a counterexample to the conjecture.) Suppose G_0 is spherically symmetric, its support is bounded, has a nonempty interior, and does not contain 0. Then, by symmetry, the Hampel-Krasker, section (i) and Krasker-Welsch solutions are of the same form. For suitable λ,

$$\psi_0(x, u) = rxh(u, \lambda/r|x|)$$

where

$$r = \left[\int |x|^2\left(2\Phi\left(\frac{\lambda}{r|x|}\right) - 1\right)G_0(dx)\right]^{m-1}.$$

If ψ_0 were a universally optimal solution for the Hampel-Krasker or MSE of prediction problems among WLS estimates, it would solve, *for all S*,

$$\text{(Vs)} \qquad \int \psi S \psi^T(x, u)\Phi(du)G_0(dx) = \min!$$

subject to $|\psi| \le \lambda$, $\psi \in \Psi$ and ψ WLS as in (3.5).

By conditioning as in the proof of Theorem 3.1 and restricting to

$$w(x, u) = \frac{\lambda}{c(x)} \frac{h(u, c(x))}{u|xR|},$$

we see that $R_0 = rI$, $c_0(x) = \lambda/r|x|$ minimizes

$$\int \lambda^2\left(\frac{d^2(xR, S)}{|xR|^2}\right)A(c(x))G_0(dx)$$

among all $c > 0$, R symmetric positive definite such that

$$\int \lambda\left(\frac{x^T xR}{|xR|}\right)B(c(x))G_0(dx) = I.$$

If we let c range over the Banach space of continuous functions vanishing at ∞ with supremum norm, it can be shown that if $p > 3$ the map

$$(c, R) \to \int \frac{x^T x R}{|xR|} B(c(x)) G_0(dx)$$

has a nonsingular differential at $c = c_0$, $R = R_0$ where r is given in the definition of ψ. Therefore by Luenberger (1969, page 243) there exists a Lagrange multiplier matrix $W_S S$ such that R_0, c_0 minimize

$$\int \frac{d^2(xR, S)}{|xR|^2} A(c(x)) G_0(dx) - 2 \int \frac{\operatorname{tr}(W_S S R x^T x)}{|xR|} B(c(x)) G_0(dx) \tag{3.24}$$

among all R symmetric positive definite, $c \geq 0$, c's vanishing at ∞. But minimization over c leads as in Theorem 3.1 to

$$c = \operatorname{tr}(RSRx^T x)/\operatorname{tr}(W_S S R x^T x) \,|\, xR \,|. \tag{3.25}$$

If we set $c = c_0$, $R = R_0$, we deduce that $W_S = R_0/\lambda$. If we now substitute (3.25) back into (3.24), find the differential of the resulting map from the set of symmetric matrices to the real line and set it equal to 0 at $R = R_0$, we obtain the equation

$$\int \alpha(c_0(x))((SR_0 + R_0 S) - 2\beta(x, S)R_0) x^T x G_0(dx) = 0 \tag{3.26}$$

where

$$\alpha(c) = 2(c\Phi(-c) - \phi(c)), \qquad \beta(x, S) = d^2(xR_0, S)/|\, xR_0 \,|^2.$$

Simplifying, we get

$$S \int \alpha\left(\frac{\lambda}{r\,|\,x\,|}\right) x^T x G_0(dx) = \int \alpha\left(\frac{\lambda}{|\,x\,|}\right) \frac{xSx^T}{|\,x\,|^2}\, x^T x G_0(dx) \tag{3.27}$$

for all positive definite symmetric S. Passing to the limit, the relationship must hold for nonnegative definite S as well. Put

$$S = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & \cdots & 0 \\ 0 & & \cdots & 0 \end{pmatrix}$$

to obtain a contradiction since by symmetry of G_0, $\int \alpha(\lambda/r\,|\,x\,|) x^T x G_0(dx)$ is a multiple of I and G_0 has a nonempty interior.

NOTES.

(1) For $p > 1$ as in the univariate case we would typically need to estimate G_0 and σ in order to implement adequate scale equivariant estimates. No new theoretical issues arise from optimality considerations. However the computational solution and existence of problems which arise with simultaneous estimation of scale become more serious.

(2) Our discussion in this section is essentially limited to the contamination neighbourhood since the maximum bias (as measured by different norms) in the p-variate case can only be easily calculated for these. However, these solutions are also adequate for variational and Kolmogorov neighbourhoods provided t is taken as double its value for contamination. Thus, for $\mathscr{F}_{\mathrm{a0v}}$, $\mathscr{F}_{\mathrm{1v}}$

$$\sup |b(\psi, F)| \le 2t \sup_{x,u} |\psi(x, u)| \tag{3.28}$$

while for $\mathscr{F}_{\mathrm{0v}}$

$$\sup |b(\psi, F)| \le 2t \int \sup_u |\psi(x, u)| G_0(dx) \tag{3.29}$$

and for $\mathscr{F}_{\mathrm{a0k}}$, $\mathscr{F}_{\mathrm{1k}}$

$$\sup_{\mathscr{F}_{\mathrm{1k}}} |b(\psi, F)| \le t \sup_x | \| \psi(x, \cdot) \| | \tag{3.30}$$

where $\| \psi(x, \cdot) \| = (\| \psi_1(x, \cdot) \|, \cdots, \| \psi_p(x, \cdot) \|)$ and $\| \psi_i(x, \cdot) \|$ is the variational norm of $\psi_i(x, \cdot)$.

(3) The invariant estimates based on minimizing MSE of prediction are appealing and seem reasonable for the error free x models. They are seriously compromised for errors in variables, however, since the matrix $\int x^T x G_0(dx)$ is not robustly estimated by replacing G_0 by the empirical distribution. A fairly artificial way out is to down weight extreme values of x. That is, let u_2 satisfy conditions of Maronna (1976), and $\Sigma(G_0)$ be the robust covariance determined by that u_2.

$$\int u_2(d(x, \Sigma^{-1})) x^T x G_0(dx) = \Sigma. \tag{3.31}$$

Then we can easily see that the estimate which minimizes the downweighted MSE of prediction

$$\sup_{\mathscr{F}} \left\{ \int u_2(d(x, \Sigma^{-1})) \{ x V(\psi) x^T + x b^T(\psi) b(\psi) x^T \} G_0(dx) \right\}$$

is given by (3.19) with Σ given by (3.31) for both $\mathscr{F}_{\mathrm{ac0}}$ and $\mathscr{F}_{\mathrm{c1}}$. The estimate is clearly equivariant. This is essentially equivalent to a proposal of Maronna, Bustos, and Yohai (1979).

APPENDIX

PROOF OF (2.11)-(2.19). For the errors in variables models these claims are proved in [B]. For the other neighbourhoods the arguments are similar. As an example here is the proof of (2.11).

Since $G = G_0$, by (1.2),

$$b(\psi, G, H) = t \int\int \psi(x, u) M(du \mid x) G_0(dx). \tag{A.1}$$

Since M is arbitrary (2.11) follows. As a second example we prove (2.17) for $\mathscr{F}_{\cdot v}$.

Write

$$
\begin{aligned}
b(\psi, G, H) &= \int\int \psi(x, u)[H(du \mid x) - \Phi(du)]G_0(dx) \\
&= \int\int \psi(x, u)[M^+(du \mid x) - M^-(du \mid x)]\alpha(x)G_0(dx)
\end{aligned}
\tag{A.2}
$$

where $\alpha(x)$ is the common total mass of the positive and negative parts of the measure $H(\cdot \mid x) - \Phi(\cdot)$ and M^+, M^- are the probability measures obtained by normalizing these positive and negative parts. $F \in \mathscr{F}_{\text{av1}}$ means $\int \alpha(x)G_0(dx) \le tn^{-1/2}$. Since M^+, M^- are arbitrary, (2.17) follows. □

PROOF OF (3.7). By definition

$$
\begin{aligned}
\mid b \mid (\psi, F) &= t\left\{\textstyle\sum_{j=1}^p\left(\displaystyle\int\int \psi_j(x, u)M(du \mid x)G_0(dx)\right)^2\right\}^{1/2} \\
&\le t \int \left\{\textstyle\sum_{j=1}^p \left(\displaystyle\int \psi_j(x, u)M(du \mid x)\right)^2\right\}^{1/2} G_0(dx)
\end{aligned}
\tag{A.3}
$$

by Jensen's inequality applied to the random vector

$$
\left(\int \psi_1(X_1, u)M(du \mid X_1), \cdots, \int \psi_p(X_1, u)M(du \mid X_1)\right).
$$

Existence of solutions in Theorem 3.1.

Sketch of argument. Consider ψ as elements of $L_2(F_0; R^p)$, square integrable p-variate functions. Define the following maps from L_2 to R or R^{p^2}

$$a_0: \psi \to \int \mid \psi \mid^2(x, u)\Phi(du)G_0(dx)$$

$$a_1: \psi \to \int \sup_u \mid \psi(x, u) \mid G_0(dx)$$

$$a_2: \psi \to \int ux^T\psi(x, u)\Phi(du)G_0(dx)$$

$$a_3: \psi \to \sup_{x,u} \mid \psi(x, u) \mid.$$

Then a_0, a_1 are convex, a_2 is linear. Let

$$\lambda_{1M} = \inf\{\lambda : \psi \in \Psi, a_1(\psi) \le \lambda, a_3(\psi) \le M\}.$$

It is easy to see that $\lambda_{1M} \downarrow \lambda_1$ if $M \to \infty$. Suppose $\lambda > \lambda_{1M}$. Then by problem 7, page 236 of Luenberger (1969) there exist Q_M, S_M such that

$$
\begin{aligned}
&\inf\{a_0(\psi): a_1(\psi) \le \lambda, a_2(\psi) = I, a_3(\psi) \le M\} \\
&\quad = \inf\{a_0(\psi) - 2 \operatorname{tr} Q[a_2(\psi) - I] + (2/s)[a_0(\psi) - \lambda]\}.
\end{aligned}
\tag{A.4}
$$

Moreover since $\{\psi: a_3(\psi) \le M\}$ is weakly compact and a_0 is lower semicontinuous, the infima in (A.4) are assumed by, say, $\psi_M^* \in \Psi$. By arguing as in the proof of the theorem

$$\psi_M^*(x, u) = \rho(x, u, s_M, Q_M) \quad \text{if} \quad |\rho(x, u, s_M, Q_M)| \le M.$$

It readily follows by considering s_M and $Q_M/\mathrm{tr}(Q_M)$ that we can extract a subsequence $\{M_r\}$ such that $\psi_{M_r}^*$ converges pointwise to a limit ψ^* as $M_r \to \infty$. Since by the optimality of $\psi_{M_r}^*$, the sequence $a_0(\psi_{M_r}^*)$ is uniformly bounded, we can conclude that $a_2(\psi_{M_r}^*) \to a_2(\psi^*)$, i.e. $\psi^* \in \Psi$ and $a_1(\psi_{M_r}^*) \to a_1(\psi^*)$. By lower semicontinuity of a_0, ψ^* is the solution to (V′). Applying (A.5) with $M = \infty$ we obtain $(s(\lambda), Q(\lambda))$ such that $\rho(x, u, Q(\lambda), s(\lambda)) = \psi^*$. Unicity of (Q, s) follows from the strict convexity of a_0. □

REFERENCES

BICKEL, P. J. (1975). One step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434.

BICKEL, P. J. (1981). Quelques aspects de la statistique robuste. In École d'Été de Probabilités de St. Flour. *Springer Lecture Notes in Math.* **876** 2–68.

HAMPEL, F. R. (1978). Optimally bounding the gross-error-sensitivity and the influence of position in factor space. *1978 Proceedings of the A.S.A. Statistical Computing Section.* A.S.A., Washington, D.C. 59–64.

HOLMES, R. (1981). Thesis, University of California, Berkeley.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–233. University of California Press.

HUBER, P. J. (1973). Robust regression: asymptotics, conjectures, and Monte Carlo. *Ann. Statist.* **1** 799–821.

HUBER, P. J. (1983). Minimax aspects of bounded influence regression. *J. Amer. Statist. Assoc.* **78** 66–80.

KRASKER, W. (1980). Estimation in linear regression models with disparate data points. *Econometrica* **48** 1333–1346.

KRASKER, W. and WELSCH, R. (1982). Efficient bounded influence regression estimation. *J. Amer. Statist. Assoc.* **77** 595–604.

LUENBERGER, D. (1969). *Optimization by Vector Space Methods.* Wiley, New York.

MARONNA, R. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67.

MARONNA, R., BUSTOS, O., and YOHAI, V. (1979). Bias and efficiency robustness of general (M) estimates for regression with random carriers. In *Smoothing Techniques for Curve Estimation* 91–116. T. Gasser and M. Rosenblatt, Eds. Springer-Verlag, Berlin.

MARONNA, R. A. and YOHAI, V. (1981). Asymptotic behaviour of general (M) estimates for regression and scale with random carriers. *Z. Wahrsch. verw. Gebeite* **58** 7–20.

RIEDER, H. (1978). A robust asymptotic testing model. *Ann. Statist.* **6** 1080–1099.

SACKS, J. and YLVISAKER, D. (1978). Linear estimation for approximately linear models. *Ann. Statist.* **6** 1122–1137.

STAHEL, W. (1981). Thesis. E.T.H. Zurich.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 93720

The Annals of Statistics
1984, Vol. 12, No. 3, 864-879

PARAMETRIC ROBUSTNESS: SMALL BIASES CAN BE WORTHWHILE[1]

BY P. J. BICKEL

University of California, Berkeley

We study estimation of the parameters of a Gaussian linear model $\mathscr{M}_0$ when we entertain the possibility that $\mathscr{M}_0$ is invalid and a larger model $\mathscr{M}_1$ should be assumed. Estimates are robust if their maximum risk over $\mathscr{M}_1$ is finite and the most robust estimate is the least squares estimate under $\mathscr{M}_1$. We apply notions of Hodges and Lehmann (1952) and Efron and Morris (1971) to obtain (biased) estimates which do well under $\mathscr{M}_0$ at a small price in robustness. Extensions to confidence intervals, simultaneous estimation of several parameters and large sample approximations applying to nested parametric models are also discussed.

1. Introduction. The basic aim of robust inference as developed by Huber, Hampel and others has been the production and study of statistical procedures which

(a) perform reasonably well when the parametric assumptions are perfectly satisfied; and

(b) are relatively insensitive to nonparametric departures from parametric assumptions which a given data set is believed to satisfy.

The main parametric model considered has been the Gaussian linear model and the departures, outliers and gross errors in the variables, have been modeled by assuming non-Gaussian error distributions and, where suitable, dependence between the independent and error variables.

An important aspect of this point of view is a focus on inference about parameters of interest rather than on deciding whether the parametric model provides an adequate fit. This is in contrast to the older approach of estimation and testing after a goodness of fit test or more generally rejection of outliers.

The same point of view makes sense in a purely parametric context. We have two possible parametric models in mind, $\mathscr{M}_0$, $\mathscr{M}_1$ with $\mathscr{M}_0 \subset \mathscr{M}_1$. Our primary interest is in estimating parameters which are identifiable in $\mathscr{M}_1$.

Again,

(i) we believe that $\mathscr{M}_0$ is adequate and want estimates or confidence regions based on estimates that perform well under that assumption. However

(ii) we wish to guard against the possible departures presented by $\mathscr{M}_1$.

Received April 1983; revised April 1984.

[1] Work performed with the partial support of Office of Naval Research Contract N00014-80-C-0163 and the Adolph and Mary Sprague Miller Foundation. Some of this material was presented at the 1980 Wald Lectures of the Institute of Mathematical Statistics.

AMS 1980 *classifications*. Primary 62F10; secondary 62F25.

Key words and phrases. Parametric robustness, pretesting, limited translation estimates, confidence intervals.

Here is the main situation we are thinking of with some specific examples.

Nested linear models. We observe $y_{n\times 1}$ where

$$y = \theta + e.$$

e is an n-variate normal vector with mean 0 and covariance matrix Σ. θ ranges freely over an r-dimensional linear space Θ_0 under $\mathscr{M}_0$ and over an s-dimensional linear space $\Theta_1 \supset \Theta_0$ under $\mathscr{M}_1$ where $r < s \leq n$. We suppose Σ known. Our asymptotic analysis in Section 5 will permit us as usual to substitute a consistent estimate $\hat{\Sigma}$ for Σ. We are interested in inference about $\mu(\theta)$ where μ is a linear function of θ. Special cases are:

1(a) *Pooling means* (Mosteller, 1948). We are given two samples $X_1, \cdots, X_m$ independent $\mathscr{N}(\mu, \sigma^2)$; $Y_1, \cdots, Y_n$ independent $\mathscr{N}(\mu + \Delta, \sigma^2)$. We want to estimate or set a confidence interval on μ. We believe $\Delta = 0$ ($\mathscr{M}_0$) but want to guard against arbitrary Δ ($\mathscr{M}_1$). Plausible examples, e.g. measurements in a current and previous survey, are discussed by Mosteller.

1(b) *Additive effects with possible interactions.* Suppose $\mathscr{M}_1$ is an ANOVA model in the sense of Scheffé (1959), possibly including random effects, which contains some interaction terms as well as main effects, and $\mathscr{M}_0$ is purely additive specifying all interactions to be 0. We take the variances of all random effects as well as measurement errors to be known. We want to study some or all of the main effects. An interesting special case is the crossover design discussed by B. W. Brown (1980). Here two groups of subjects I and II which for simplicity we take of equal size $n/2$ are each administered two drugs A, B in succesion and responses measured. The second drug is administered after response to the first has been measured and a time deemed sufficient for the effect of the first to wear off has elapsed. The order of administration of the drugs is AB in group 1, BA in group 2. Model $\mathscr{M}_1$ here is that the response $Y_{ijk(u)}$ of the jth subject in group i during period k who is administered drug u during that period is

$$Y_{ijk(u)} = \mu + \pi_k + \phi_u + \lambda_{uk} + \xi_{ij} + \varepsilon_{ijk}$$

where π_k, $k = 1, 2$, is the period effect, ϕ_u, $u =$ A, B is the drug effect, and λ_{uk} is the interaction of drug u and period k with $\lambda_{u1} = 0$. These are all fixed. As usual, identifiability requires further linear restrictions. On the other hand, ξ_{ij}, the effect of the jth subject in group i, is considered random $\mathscr{N}(0, \sigma_\xi)$, and ε_{ijk}, the within subject deviation for the kth period (including measurement error), is modeled as $\mathscr{N}(0, \sigma_\varepsilon^2)$. All are modeled as independent of each other. We assume σ_ξ^2, σ_ε^2 known. $\mathscr{M}_0$ specifies that, as we hope, there is no interaction, $\lambda_{uk} \equiv 0$. We are interested in estimating $\phi_b - \phi_a$, the difference in effectiveness of the drugs.

1(c) *Nested regression models.* Write $\theta = X\beta$, $\beta_{s\times 1}$, $X = (x_1, \cdots, x_s)$ an $n \times s$ matrix of rank s and think of the s columns of X as corresponding to s independent variables. Suppose β ranges freely over R^s under $\mathscr{M}_1$ but $s - r$ coordinates of β are set equal to 0 under $\mathscr{M}_0$, i.e. $s - r$ of the independent variables are irrelevant. Various linear functions $\mu(\theta)$ are of interest, for instance the vector of expectations θ itself or one or more predicted values $x\beta$, at various values x.

From this special case we will proceed (under regularity conditions) by an asymptotic analysis to the general case of

Nested parametric models. We observe $(X_1, \cdots, X_n)$ with joint density $p_n(x, \theta)$ (with respect to some measure ν_n). Under $\mathscr{M}_1$, $\theta \in \Theta_1$, an open subset of s-dimensional space. Under $\mathscr{M}_0$, $\theta \in \Theta_0 \subset \Theta_1$, a (locally) r-dimensional subsurface of Θ_1, and μ is a smooth vector-valued function of θ. This of course covers all previous situations as well as many others including Example 1 with σ^2 unknown, nested loglinear models, etc.

Our point of view, essentially already suggested by Hodges and Lehmann (1952), page 402, is that procedures should be judged by their maximum risks under $\mathscr{M}_0$ and $\mathscr{M}_1$. So, in the context of nested parametric models, if $M(\theta, \delta)$ is the risk of a decision rule δ when θ is true we should look at

$$m(\delta) = \sup\{M(\theta, \delta): \theta \in \Theta_0\}, \quad M(\delta) = \sup\{M(\theta, \delta): \theta \in \Theta_1\}.$$

M can be thought of as a measure of robustness of δ and we should be interested in procedures which make m small subject to a bound on M.

In the basic linear model example the solutions we end up with are necessarily biased under $\mathscr{M}_1$. Robustness requires that the biases be bounded through M. The worthwhile gains are in reduction of m over the unbiased minimax estimate.

In Section 2 we apply this theory to the linear model example for quadratic loss when μ is one dimensional. The optimal procedures are difficult to compute. We motivate a family of reasonable approximately optimal solutions, compare them numerically to the optimum and other competitors and also briefly discuss the crucial question of selection within the family.

In Section 3 we discuss confidence intervals based on these estimates. In Section 4, we derive, using results of Berger (1982) and Huber (1977), some procedures for the multivariate case. In Section 5, we show how these ideas generalize to yield reasonable procedures in nested parametric models and, finally, in Section 6, give conclusions and propose open questions.

2. The nested linear models: dim(μ) = 1, quadratic loss.

a) *Optimality theory.* We specialize to estimation of μ with quadratic loss. That is, we assume that μ is real, linear, and if $\delta(x)$ is an estimate

$$M(\theta, \delta) = E_\theta(\delta(X) - \mu(\theta))^2. \tag{2.1}$$

Since we assume Σ known, we can, by taking $Y^* = Y\Sigma^{-1/2}$, $\mathscr{M}_i^* = \mathscr{M}_i\Sigma^{-1/2}$, reduce our problem to one in which the observation Y^* has covariance matrix $\sigma^2 I$, the standard linear model.

Let $\hat{\mu}_i = \mu(\hat{\theta}_i)$, $i = 0, 1$, be the least squares estimates of μ under $\mathscr{M}_0$, $\mathscr{M}_1$ respectively. Then, for $i = 0, 1$, $\hat{\mu}_i$ has constant risk and is minmax under $\mathscr{M}_i$. Let

σ_i^2 be the variance of $\hat{\mu}_i$ so that

$$\inf_\delta M(\delta) = \sigma_1^2, \quad \inf_\delta m(\delta) = \sigma_0^2.$$

Let $\hat{\mu}_c^*$ minimize $m(\delta)$ subject to $M(\delta)/\sigma_1^2 \leq 1/c$ so that $\hat{\mu}_i^* = \hat{\mu}_i$, $i = 0, 1$. Note that $M(\hat{\mu}_0) = \infty$ and $\hat{\mu}_0$ is certainly not robust. Let

$$\rho = \operatorname{corr}(\hat{\mu}_0, \hat{\mu}_1) = \sigma_0/\sigma_1 \tag{2.2}$$

which is independent of the error variance σ^2,

$$\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0 \tag{2.3}$$

and

$$\sigma_{\hat{\Delta}}^2 = \sigma_1^2(1 - \rho^2), \tag{2.4}$$

its variance.

PROPOSITION 1. *The estimate $\hat{\mu}_c^*$ may be written*

$$\hat{\mu}_c^* = \hat{\mu}_0 + \sigma_{\hat{\Delta}} w_q^*(\hat{\Delta}/\sigma_{\hat{\Delta}}) \tag{2.5}$$

where

$$q^2 = (1 - c)/c(1 - \rho^2) \tag{2.6}$$

(2.7) *w_q^* is odd and obtained by minimizing $Ew^2(Z)$ subject to $\sup_\Delta E(w(Z + \Delta) - \Delta)^2 \leq 1 + q^2$ for $Z \sim \mathscr{N}(0, 1)$.*

NOTE. Evidently w_q^* is the solution of the special case $\mu = \theta$, $r = 0$, $s = 1$, $\sigma^2 = 1$. We call this problem (P).

PROOF. By sufficiency reduce to $\hat{\theta}_1$ and without loss of generality choose a canonical basis so that $\hat{\theta}_0$ consists of the first r components of $\hat{\theta}_1$ and all components of $\hat{\theta}_1$ are independent normal variables with variance σ^2. Moreover we can arrange that $\hat{\mu}_0/\sigma_0$ is the first component of $\hat{\theta}_1$ and $\hat{\Delta}/\sigma_1(1 - \rho^2)^{1/2}$ is the $(r+1)$st component. Note by Hodges and Lehmann (1952) that $\hat{\mu}_c^*$ is unrestrictedly minimax for the "mixed" model: for suitable $\lambda(c)$ and $\theta = (\theta^{(1)}, \cdots, \theta^{(s)})$, $\hat{\theta}_1$ has density $(1 - \lambda)p_1 + \lambda p_0$ where p_1 is the density of $\hat{\theta}_1$ under $\mathscr{M}_1$ and θ, while p_0 is the density of $\hat{\theta}_1$ under $(\theta^{(1)}, \cdots, \theta^{(r)}, 0, \cdots, 0)$, i.e. under $\mathscr{M}_0$. We can reduce this unrestricted problem by invariance, using for instance Kiefer's (1957) general results. Since we want to estimate

$$\sigma_0\theta^{(1)} + (1 - \rho^2)^{1/2}\sigma_1\theta^{(r+1)},$$

the problem is invariant under arbitrary translations of $\theta^{(i)}$, $i \neq 1, r + 1$, and we can reduce to $\hat{\mu}_0$, $\hat{\Delta}$. The problem is also invariant under translations of $\hat{\mu}_0$, keeping $\hat{\Delta}$ fixed. Since $\hat{\mu}_c^*$ is unique it therefore must be of the form $\mu_0 + w(\hat{\Delta})$. Claims (2.7) and (2.6) follow by calculation. □

Unfortunately calculation of w_q^* is difficult. See Bickel (1983) for its rather unpleasant qualitative features.

In view of these unpleasant features, it is natural to seek other families of robust estimates with more satisfactory behaviour. By invariance it seems reasonable to look for $\hat{\mu}$ of the form

$$\hat{\mu}_0 + \sigma_{\hat{\Delta}} w(\hat{\Delta}/\sigma_{\hat{\Delta}}). \tag{2.8}$$

For any such estimate

$$M(\hat{\mu}) = \sigma_1^2(\rho^2 + (1 - \rho^2)\sup_\Delta E(w(Z + \Delta) - \Delta)^2) \tag{2.9}$$

$$m(\hat{\mu}) = \sigma_1^2(\rho^2 + (1 - \rho^2)Ew^2(Z)). \tag{2.10}$$

Abusing notation, let us call the coefficients of $(1 - \rho^2)$ inside parentheses in these expressions $M_0(w)$, $m_0(w)$. They correspond to M and m in problem (P).

b) *"Approximate" optimality in problem* (P). From (2.9) and (2.10) reasonable w in problem (P) correspond to reasonable $\hat{\mu}$. In problem (P) we observe $X = Z + \Delta$, $Z \sim \mathcal{N}(0, 1)$ and we want to minimize $m_0(w)$ subject to a bound on $M_0(w)$. Three approximate optimality principles lead to the same family, the limited translation estimates of Efron and Morris (1971) defined by

$$\begin{aligned} e_q(x) &= 0, & |x| \le q \\ &= x - q \operatorname{sgn} x, & |x| > q, \end{aligned}$$

which leads to $M_0(e_q) = 1 + q^2$.

I. *Optimality in a related problem* (Bickel, 1983, Marazzi, 1980). Suppose π is a prior distribution, $r(\pi)$ the Bayes risk, w_π the Bayes estimate, and $G_\pi = \pi * \Phi$, where $*$ denotes convolution, is the marginal distribution of X. Then,

$$r(\pi) = 1 - I(G_\pi) \tag{2.11}$$

$$w_\pi(x) = x + (g'_\pi/g_\pi)(x) \tag{2.12}$$

where g_π is the density of G_π, $I(G)$ is the Fisher information where

$$\begin{aligned} I(G) &= \int \frac{[g']^2}{g}(x)\,dx, && \text{if the integral is defined} \\ &= \infty && \text{otherwise.} \end{aligned}$$

By Hodges and Lehmann (1952) and (2.11), the optimal w_q^* corresponds to G_q^* which for some $\lambda(q)$ minimizes $I(G)$ over $\mathcal{G}_0 = \{G = (1 - \lambda)\Phi + \lambda\Phi * H,\ H$ arbitrary$\}$. If we "approximate" $\mathcal{G}_0$ by $\mathcal{G}_1 = \{G = (1 - \lambda)\Phi + \lambda H,\ H$ arbitrary$\}$ we arrive at Huber's (1964) problem with solution G_1 where

$$\begin{aligned} (g_1'/g_1)(x) &= -x, & |x| \le q \\ &= -q \operatorname{sgn} x, & |x| > q. \end{aligned}$$

Substituting into (2.12), we get the Efron-Morris family.

II. *Bounding unbiased estimate of risk* (Berger, 1982). If

$$\psi(x) = x - w(x) \tag{2.13}$$

under mild conditions

$$M(\Delta, w) = 1 + E_\Delta(\psi^2(x) - 2\psi'(x))$$

so that $1 + \psi^2(x) - 2\psi'(x)$ is the UMVU estimate of $M(\eta, w)$. Berger (in a more general context) proposes minimizing $m_0(w)$ subject to $\psi^2(x) - 2\psi'(x) \le q^2$. The solution is easily seen to be e_q.

In fact Berger's approach must yield the same results as approach I both in our context and his more general restricted Bayes models. To see this in our model, note that

$$\begin{aligned}
&\inf_w\{(1-\lambda)m_0(w) + \lambda \sup_x(1 + \psi^2(x) - 2\psi'(x))\} \\
&= 1 + \inf_\psi \sup\left\{\int (\psi^2(x) - 2\psi'(x))G(dx) \colon G \in \mathscr{G}_1\right\} \\
&= 1 - \min\{I(G) \colon G \in \mathscr{G}_1\}
\end{aligned}$$

by a minmax argument.

III. *Bounding unbiased estimate of bias.* Note that $\psi(X)$ is the UMVU estimate of the bias of $w(X)$. Thus it seems reasonable to minimize $m_0(w)$ subject to $\sup_x |\psi(x)| \le q$. This is the exact analogue of Hampel's robustness formulation. The solution is again e_q.

For further optimality properties of Efron-Morris estimates, see Bickel (1983).

c) *Performance of Efron-Morris* (E-M) *estimates and competitors.* We measure the relative performance of estimates $\hat{\mu}$ by their relative savings and losses in risk with respect to $\hat{\mu}_1$

$$S(\hat{\mu}) = 1 - m(\hat{\mu})/m(\hat{\mu}_1), \quad L(\hat{\mu}) = M(\hat{\mu})/M(\hat{\mu}_1) - 1.$$

For estimates of the form (2.8),

$$S(\hat{\mu}) = (1 - \rho^2)(1 - m_0(w)), \quad L(\hat{\mu}) = (1 - \rho^2)(M_0(w) - 1).$$

Table 1 gives $1 - m_0(w)$ as a function of $q^2 = M_0(w) - 1$ for the E-M estimates, for w_q^* (calculated by Dr. A. Marazzi) and for some competitors which we now discuss.

Pretesting estimates. A type of procedure long advocated by Bancroft and others (see Bancroft and Han, 1977, for a review) are estimates

$$\begin{aligned}
\hat{\mu} &= \hat{\mu}_0, \quad |\hat{\theta}_1 - \hat{\theta}_0| \le c\sigma \\
&= \hat{\mu}_1, \quad \text{otherwise}
\end{aligned}$$

with c chosen to produce an appropriate level for the test of $H\colon \mathscr{M}_0$ vs. $\mathscr{M}_1$ based

TABLE 1
Gain at 0, $g = 1 - m_0(\omega)$, as a function of the increase in maximum risk $q^2 = M_0(w) - 1$.

q^2	g_e	g_b	g_s	g_j	q	$d(q)$
.1	.413	.085	—	—	.316	.715
.2	.538	.155	—	.330	.447	.903
.3	.619	.225	—	.438	.548	1.053
.4	.676	.290	—	.523	.632	1.175
.5	.721	.350	.711	.592	.707	1.281
.6	.758	.405	.753	.648	.775	1.370
.7	.786	.455	.788	.695	.837	1.461
.8	.811	.500	.816	.735	.894	1.538
.9	.832	.540	.840	.768	.949	1.608
1.0	.850	.58	.859	.796	1.000	1.679

Note: g_e is the increase for the E-M estimate, g_b for the pretest, g_s for the Sacks family, g_j for Jeffreys' type of generalized Bayes estimate. q and $d(q)$ are the critical values for the E-M and pretest estimates.

on $(|\hat\theta_1 - \hat\theta_0|)/\sigma$. If $|\hat\mu_1 - \hat\mu_0| \neq |\hat\theta_1 - \hat\theta_0|$, this estimate is not of the form (2.8). A version of that form can be based on testing H: $E\hat\Delta = 0$ vs. $E\hat\Delta \neq 0$ and is given by

$$\hat\mu_c^B = \hat\mu_0 + \sigma_{\hat\Delta} b_q\left(\frac{\hat\Delta}{\sigma\hat\Delta}\right)$$

with

$$b_q(x) = 0, \quad |x| \le d(q) \qquad (2.14)$$
$$= x, \quad |x| > d(q)$$

and d chosen so that

$$M_0(b_q) = 1 + q^2.$$

The ψ function corresponding to b_q via (2.13) corresponds to hard rejection which is known not to work well. This seems true here too. The Bancroft-Han estimate is even worse. (See also Sclove et al. (1972).

Another interesting and desirable feature of the E-M family is monotonicity of $M(\Delta, e_q)$ as a function of $|\Delta|$, i.e. $M_0(e_q)$ is assumed at $|\Delta| = \infty$. This is not true of the pretest estimates and more generally estimates which correspond to redescending ψ functions. Nevertheless we can expect smooth versions of such estimates to perform reasonably well. Motivated by Sacks and Ylvisaker (1978), J. Sacks has proposed a family of such ψ,

$$\psi_\gamma(x) = 2(2 + (|x| - \gamma)_+^2)^{-1}x.$$

Another natural family consists of the Jeffreys' type estimates which are generalized Bayes with respect to a prior distribution placing mass p at 0 and corresponding to Lebesgue measure otherwise.

$$\delta_p(x) = x((1/p - 1)\varphi(x) + 1)^{-1}.$$

Table 1 shows very substantial gains in m_0 for small payments in M_0. Small biases can be very worthwhile. The pretest estimates are clearly poor and the Jeffreys type estimates are inferior to both the E-M and Sacks estimates.

There is, of course, a serious question as to which E-M estimate to use. The natural way is to calibrate by the maximum $L(\hat{\mu})$ we are willing to tolerate. This of course depends both on ρ^2 and $M_0(w)$. For instance, if $n_1 = n_2$ in the pooling example $\rho^2 = 1/2$. If we are willing to accept a 10% loss we would take $q = .2$ and obtain a gain of (.5) (.538) = 26.9%.

Another idea is to bound the maximum squared bias of $\hat{\mu}$ standardized by the variance of $\hat{\mu}_1$. For the E-M estimates this equals $L(\hat{\mu})$. The remaining approach of choosing d according to a reasonable level for the test of H: $\Delta = 0$ based on $\hat{\Delta}$ yields unreasonably high values of $L(q)$ and is not recommended.

The performance of E-M is markedly better than that of the "Jeffreys" or pretest procedures for small q^2. This is in accordance with the asymptotic results of Bickel (1983). Since the Sacks' procedures which are on the whole comparable with E-M cannot be extended over the whole q^2 range, we are left with E-M as the candidate of choice.

The best we can do in terms of $m_0(w)$ for given $M_0(w)$ cannot be calculated exactly. However effective numerical procedures have been derived in Marazzi (1980, 1982). Here is a table of the optimal g based on results he has supplied.

q	.06	.12	.19	.29	.44	.70
g_0	.39	.49	.57	.66	.74	.82

3. Nested linear models: μ univariate.

Confidence intervals and other loss functions. In univariate estimation problems, we usually want confidence intervals as well as point estimates. Since, given our assumed knowledge of σ, we can form fixed width confidence intervals based on $\hat{\mu}_1$, it seems reasonable to ask how intervals of the same width based on estimates $\hat{\mu}$ perform. This boils down to fixing a width $2z\sigma_1$ and using the loss function

$$\ell(\theta, d) = 1 \text{ if } |d - \mu(\theta)| \geq z\sigma_1 \qquad (3.1)$$
$$= 0 \text{ otherwise}$$

$$M(\theta, \hat{\mu}) = P[|\hat{\mu} - \mu(\theta)| \geq z\sigma_1] = 1 - P_\theta[\mu(\theta) \in \hat{\mu} + z\sigma_1]. \qquad (3.2)$$

From the argument of Proposition 1 it is easy to see that for any loss function of the form $\ell(|\mu(\theta) - d|)$, equivariant estimates are of the form (2.8). Calculation of the optimal procedures is even more hopeless for this loss function. However, it is easy to see that approximate optimality approach III continues to yield the E-M estimate. More generally

PROPOSITION 2. *Suppose $\ell(\theta, d) = \ell(|\mu(\theta) - d|)$ and ℓ is nondecreasing. Then $m(\hat{\mu})$ is minimized among all equivariant $\hat{\mu}$ of the form* (2.8) *with $|\psi(x)| \leq q$ by an* E-M *estimate*

$$\hat{\mu}_c^e = \hat{\mu}_0 + \sigma_{\hat{\Delta}} e_q(\hat{\Delta}/\sigma_{\hat{\Delta}}). \qquad (3.3)$$

PROOF. Without loss of generality, suppose $\sigma_{\hat{\Delta}} = 1$. If $\theta \in \Theta_0$ and $\hat{\mu}$ is given by (2.8)

$$m(\hat{\mu}) = E\ell(|U + w(V)|)$$

where U, V are independent normal with mean 0. By Anderson's theorem (Anderson, 1955) $E(\tilde{\ell}\,(|U + w(V)|)\,|\,V)$ is monotone increasing in $|w(V)|$. The proposition follows. □

The risk of an E-M estimate (3.3) for a loss function $\ell(|\theta - d|)$ is given by

$$M(\theta, \hat{\mu}_c^e) = \int_{-\infty}^{\infty} \Big\{ \ell(\sigma_0 u - \Delta)[\Phi(d - \tilde{\Delta}) - \Phi(-q - \tilde{\Delta})]$$

$$+ \int_{q-\tilde{\Delta}}^{\infty} \ell(\sigma_0 u + \sigma_1(1 - \rho^2)^{1/2}(w - q))\phi(w)\, dw \tag{3.4}$$

$$+ \int_{-\infty}^{-q-\tilde{\Delta}} \ell(\sigma_0 u + \sigma_1(1 - \rho^2)^{1/2}(w + q))\phi(w)\, dw \Big\} \phi(u)\, du$$

where $\Delta = \mu(\theta) - \mu(\theta_0)$, $\tilde{\Delta} = \Delta/\sigma_1(1 - \rho^2)^{1/2}$. Evidently M depends on θ through Δ only, as it must, and moreover,

PROPOSITION 3. *If ℓ is as in Proposition 2, then M is a nondecreasing function of $|\Delta|$ for the estimator $\hat{\mu}_c^e$.*

PROOF. It is enough to consider ℓ such that ℓ' exists and is bounded since we can then obtain the general case by approximation. Differentiate M with respect to Δ and interchange limits to get

$$\frac{\partial M}{\partial \tilde{\Delta}}(\theta, \hat{\mu}_c^e)$$
$$= \sigma_1(1 - \rho^2)^{1/2}[\Phi(q - \tilde{\Delta}) - \Phi(-q - \tilde{\Delta})] \int_{-\infty}^{\infty} \tilde{\ell}\,'(\sigma_0 u - \Delta)\phi(u)\, du \geq 0. \quad □$$

NOTE. This establishes monotonicity of risk for an arbitrary monotone loss function in the original problem considered by Efron and Morris. Thus

$$m(\hat{\mu}_c^e) = \left(\int_{-\infty}^{\infty} \ell(\sigma_0 u)\phi(u)\, du\right)(2\Phi(q) - 1)$$
$$+ 2\int_{-\infty}^{\infty}\int_{d}^{\infty} \ell(\sigma_0 u + \sigma_1(1 - \rho^2)^{1/2}(v - q))\phi(v)\phi(u)\, du\, dv \tag{3.5}$$

$$M(\hat{\mu}_c^e) = \int_{-\infty}^{\infty} \ell(\sigma_1(u - (1 - \rho^2)^{1/2}q))\phi(u)\, du. \tag{3.6}$$

TABLE 2
Minimum probabilities of coverage of fixed length intervals centered at E-M estimates: z = 1.960.

q^2	.2	.4	.6	.8
.2	.982	.978	.972	.962
	.932	.936	.941	.945
.4	.988	.985	.977	.965
	.912	.922	.932	.941
.6	.992	.989	.980	.966
	.894	.908	.922	.936
.8	.994	.991	.982	.968
	.874	.894	.913	.932

Note: For each table, the first entry in each box is the minimum probability of coverage on $\mathscr{M}_0$ given by (3.7), the second the minimum on $\mathscr{M}_1$ given by (3.8).

If we specialize to confidence intervals as in (3.1), we obtained minimum probabilities of coverage,

$$
\begin{aligned}
1 - m(\hat{\mu}_c^e) &= (2\Phi(z/\rho) - 1)(2\Phi(q) - 1) \\
&\quad + 2P[-z - (1-\rho^2)^{1/2}q \le A \le z - (1-\rho^2)^{1/2}d,\ B \ge q]
\end{aligned} \tag{3.7}
$$

where (A, B) are bivariate standard normal with correlation $(1-\rho^2)^{1/2}$.

$$1 - M(\hat{\mu}_c^e) = \Phi(z - (1-\rho^2)^{1/2}q) + \Phi(z + (1-\rho^2)^{1/2}q) - 1. \tag{3.7a}$$

We give these probabilities for $z = 1.96$ (corresponding to a 95% confidence level) and selected q in Table 2. The results are similar for the 90% and 99% levels. Again the cost benefit structure seems attractive.

Brown (1980) essentially uses pretest estimate based confidence intervals on a data set to illustrate the dangers of the crossover method. If we treat σ_ξ^2, σ_ϵ^2 as equal to their estimated values so that $\rho^2 = .48$ for these data and say select $q = .2$ in Table 1 so that $L(\hat{\mu}_q^e) \cong .10$ we obtain significant results for all ($\mathscr{M}_1$) confidence levels tabled and a fortiori all corresponding ($\mathscr{M}_0$) levels, which is consistent with an analysis of the data based on first period results only.

4. Nested linear models: Quadratic loss in the multivariate case. Suppose $\dim(\mu) = p$. Then $\hat{\mu}_1 \sim \mathscr{N}_p(\mu(\theta), \Sigma_1)$, $\hat{\mu}_0 \sim \mathscr{N}_p(\mu(\theta_0), \Sigma_0)$ where θ_0 is the projection of θ on Θ_0. If $\ell(\theta, d)$ is a function of $\mu(\theta) - d$, invariance considerations lead as before to estimates

$$\hat{\mu} = \hat{\mu}_0 + w(\hat{\Delta}) \tag{4.1}$$

where $\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0$ is independent of $\hat{\mu}_0$ with an $\mathscr{N}_p(\Delta, \Sigma_1 - \Sigma_0)$ distribution, $\Delta = \mu(\theta) - \mu(\theta_0)$. Specialize further to,

$$\ell(\mu(\theta) - d) = (\mu(\theta) - d)A(\mu(\theta) - d)^T, \quad A \text{ positive definite.}$$

Then,

$$m(\hat{\mu}) = \mathrm{tr}(A\Sigma_0) + \mathrm{tr}(AE_0(w^T w(\hat{\Delta})))$$

$$M(\hat{\mu}) = \mathrm{tr}(A\Sigma_0) + \sup_\Delta \mathrm{tr}(AE_\Delta((w(\hat{\Delta}) - \Delta)^T(w(\hat{\Delta}) - \Delta)))$$

and in minimizing $m(\hat{\mu})$ subject to a bound on M we need only consider the second terms above. That is, it is enough to consider the special case $r = 0$, $s = p$. Exact solution is impossible. However we can attempt approximations. We can always reduce to the case $A = \| a_i^2 \delta_{ij} \|$ diagonal, $\Sigma_1 - \Sigma_0$ the identity. That is, we observe $X = \Delta + Z$, $Z \sim \mathcal{N}_p(0, I)$, $\Delta = (\Delta_1, \cdots, \Delta_p)$. The risk of an estimate $w = (w_1, \cdots, w_p) = x - \Psi(x)$ is

$$M(\Delta, w) = \textstyle\sum_{i=1}^p a_i^2 E(w_i(X) - \Delta_i)^2$$
$$= \textstyle\sum_{i=1}^p a_i^2 + E\left\{\sum_{i=1}^p a_i^2(\psi_i^2(X) - 2\,\frac{\partial \psi_i}{\partial x_i}\,(X))\right\}$$

under mild conditions. If π is a Bayes prior distribution with Bayes risk $r(\pi)$, Bayes estimate w_π, and marginal density g_π then

$$w_\pi(x) = x + \nabla \log g_\pi(x)$$
$$r(\pi) = \textstyle\sum_{i=1}^p a_i^2 - I(G_\pi) \tag{4.2}$$

where ∇ is the gradient $((\partial/\partial x_1), \cdots, (\partial/\partial x_p))$

$$I(G) = \sum a_i^2 \int \left(\frac{\partial g}{\partial x_i}\,(x)\right)^2 g^{-1}(x)\, dx \tag{4.4}$$

(and $= \infty$ if the quantity on the right is undefined). Again the original problem is to minimize $I(G)$ over $\mathscr{G}_0$ and approximation (I) is to minimize over $\mathscr{G}_1$ (with Φ now the p-variate standard normal). By the argument given for one dimension, this yields the same solution as does approximation (II) which minimizes $M(0, w)$ subject to a bound on $[\sum a_i^2(\psi_i^2(x) - 2\,(\partial\psi_i/\partial x_i)(x))] \le q^2$, for suitable q^2. Unfortunately this approximation is also difficult to compute (but see Chen, 1983), unless all the a_i^2 are equal, say to $1/p$. In this case the solution is given for $p = 3$ by Huber (1977) and for general p by Berger (1981), Theorem 3. Here

$$\begin{aligned} w(x) &= 0 & |x| \le q \\ &= \rho(|x|^2)x, & |x| > q \end{aligned} \tag{4.5}$$

with ρ a ratio of Bessel functions with parameters depending on p and scale depending on q^2 and $\rho(|q|^2) = 0$. For $p \ge 3$ we can take $q = 0$, i.e., find the minimax estimate in this class which minimizes $M(0, w)$. The answer is the Stein positive part estimate, $q^2 = 2(p - 2)$,

$$\rho(r) = \left(1 - \frac{2(p-2)}{r}\right).$$

As Berger points out, $M(0, w)$ for this estimate drops very sharply from .296 when $p = 3$ to .07 for $p = 5$. Although ths solution is appealing we face the usual ambiguities of the multivariate case. For $p \geq 3$ we could, for instance, also reduce $M(\theta, \hat{\mu})$ for $|\mu(\theta_0)|$ small by applying Steinian shrinking to $\hat{\mu}_0$. Moreover, the effect of the choice of loss function on the suitability of the estimate is difficult to make precise.

For $a_i^2 = 1/p$, it seems reasonable to consider average squared bias and,

$$\text{minimize } E\{\textstyle\sum_{i=1}^p w_i^2(X)\} \text{ subject to } p^{-1} \textstyle\sum_{i=1}^p \psi_i^2 \leq q^2.$$

The solution is as in the one-dimensional case,

$$\begin{aligned} \tilde{w}(x) &= 0, & |x|^2 &\leq q^2 \\ &= (1 - (q/|x|)x, & |x|^2 &> q^2. \end{aligned} \tag{4.6}$$

If we define M as in the introduction then for fixed $M(w) = 1 + q^2$, estimate (4.5) improves (4.6) at $\Delta = 0$. This follows since the estimates (4.6) also have, if $\tilde{\psi}$ corresponds to $\tilde{w}$,

$$M(\tilde{w}) = 1 + p^{-1} \sup_x \sum \left[\tilde{\psi}_i^2(x) - 2\frac{\partial \tilde{\psi}_i}{\partial x_i}(x)\right] = 1 + q^2. \tag{4.7}$$

The difference is substantial and despite its attractive feature of computability for more general loss functions, this analogue to Hampel robustness seems unsatisfactory for this application.

5. Nested parametric models: Asymptotics. We extend the approaches of Sections 3 and 4 to general nested parametric models by using large sample approximations. Related results are given by Sen (1979) for pretesting estimates. For simplicity we consider estimation of $\mu(\theta)$ where μ is a smooth real-valued function of θ.

Suppose Θ_1, Θ_0 are as we described previously, respectively an open subset of R^s and a (locally) r-dimensional submanifold of Θ_1. Suppose that the models are approximable locally in the sense of Le Cam, to scale $n^{-1/2}$, by nested Gaussian linear models and admit estimates $\hat{\theta}_{0n}$, $\hat{\theta}_{1n}$ (typically M.L.E.'s under $\mathscr{M}_0$, $\mathscr{M}_1$) which are efficient and locally sufficient uniformly on compact subsets of Θ_0, Θ_1 respectively. See Le Cam (1969), Chapters 3, 4 for a detailed description of these concepts and suitable conditions.

Fix $\theta_0 \in \Theta_0$ and reparametrize Θ by $\theta_0 + an^{-1/2}$ in Pitman form. Locally Θ permits arbitrary a while Θ_0 specifies $a \in V(\theta_0)$ an r-dimensional subspace of R^s. Also $\mu(\theta_0 + an^{-1/2}) = \mu(\hat{\theta}_0) + a\dot{\mu}(\theta_0) + O(n^{-1/2})$ where $\dot{\mu}$ is the differential of μ. Finally, $n^{1/2}\{(\hat{\theta}_{0n} - \theta_0), (\hat{\theta}_{1n} - \theta_0)\}$ is asymptotically normal uniformly on compact sets of (θ_0, a) with means $(a\Pi(\theta_0), a)$ and covariance matrix $\Sigma(\theta_0)$ where $\Pi(\theta_0)$ is the projection matrix of $V(\theta_0)$.

These approximations suggest that in order to minimize maximum M.S.E. of estimates of $\mu(\theta)$ over large Pitman neighbourhoods of θ_0 in Θ_0, subject to a bound on the maximum M.S.E. over large Pitman neighbourhoods of θ_0 in Θ, we use asymptotically equivariant estimates as follows. Let

$$\hat{\Delta}_n = \mu(\hat{\theta}_{1n}) - \mu(\hat{\theta}_{0n}), \quad \sigma_\Delta^2(\theta_0) = \dot{\mu}^T(\theta_0)\begin{pmatrix}-1\\1\end{pmatrix}\Sigma(\theta_0)\begin{pmatrix}-1\\1\end{pmatrix}^T\dot{\mu}(\theta_0)$$

denote the asymptotic variance of $n^{1/2}\hat{\Delta}_n$ under $\theta_0 + an^{-1/2}$,

$$\Delta = a(I - \Pi(\theta_0))\dot{\mu}(\theta_0)$$

denote its asymptotic mean, and $\hat{\sigma}_{\Delta,n}$ be a consistent estimate of σ_Δ, e.g.

$$\hat{\sigma}_{\Delta_n} = \sigma_\Delta(\hat{\theta}_{1n}).$$

Then, an asymptotically equivariant estimate is one of the form

$$\mu(\hat{\theta}_{0n}) + \hat{\sigma}_{\Delta_n} w(\hat{\Delta}_n/\hat{\sigma}_{\Delta_n}) \tag{5.1}$$

and n times the M.S.E. at $\theta_0 + an^{-1/2}$ of such an estimate is (under mild conditions) approximated by

$$M(\theta_0, a, w) = \sigma_1^2(\theta_0)(\rho^2(\theta_0) + (1 - \rho^2(\theta_0))E(w(Z + \Delta) - \Delta)^2) \tag{5.2}$$

where $\sigma_i^2(\theta_0)$ is the asymptotic variance of $n^{1/2}\mu(\hat{\theta}_{in})$ and,

$$\rho^2(\theta_0) = \sigma_0^2(\theta_0)/\sigma_1^2(\theta_0).$$

From (5.2), given a bound $1/c$ on $\sup_a M(\theta_0, a, w)/\sigma_1^2(\theta_0)$, we minimize $\sup_{a\in V(\theta_0)} M(\theta_0, a, w)$ by taking $w = w_q^*$. As in Section 2, we obtain reasonable results by taking $w = e_q$, with q related to c via (2.6) and $\rho = \rho(\sigma_0)$. The asymptotic sufficiency and efficiency properties of $\hat{\theta}_{in}$, $i = 0, 1$, enable us to formulate asymptotic optimality and near optimality properties of these estimates in the class of all estimates. For simplicity, we omit these.

We give a simple illustration of this approach by applying it to the case of nested linear models with $\Sigma = \sigma^2 I$, σ^2 unknown, and μ a linear function of the mean θ. Then our prescription is merely to replace σ_Δ^2 in (2.8) by

$$\hat{\sigma}_\Delta^2 = \tau^2[\sigma^{-2}(\sigma_1^2 - \sigma_0^2)] \tag{5.2a}$$

where $\tau^2 = \| Y - \hat{\theta}_1 \|^2/(n - 2)$, the usual estimate of σ^2. The ratio in parentheses in (5.2) depends on the models only. For general Σ, given a consistent estimate $\hat{\Sigma}$ of Σ, we can calculate $\hat{\theta}_0$, $\hat{\theta}_1$ by generalized least squares using $\hat{\Sigma}$ and then plug $\hat{\Sigma}$ into σ_Δ^2 appropriately calculated.

As a second illustration, consider pooling two binomial samples. Let $\hat{p}_i = N_i/n_i$, $i = 1, 2$, where N_i is $\text{bin}(n_i, p_i)$, $0 < p_i < 1$, $n_1/n_2 = \lambda$, $0 < \lambda < 1$. We want to estimate p_1. $\mathscr{M}_0$ prescribes $p_1 = p_2$. So, if we use $n = n_1 + n_2$ as an index,

$$\hat{\theta}_{1n} = (\hat{p}_1, \hat{p}_2), \quad \hat{\theta}_{0n} = (\hat{p}, \hat{p})$$

where

$$\hat{p} = (N_1 + N_2)/n = (\lambda\hat{p}_1 + \hat{p}_2)/(1 + \lambda).$$

If $\theta = (p, p)$,

$$\sigma_0^2(\theta) = p(1 - p), \quad \sigma_1^2(\theta) = p(1 - p)\,\frac{(1 + \lambda)}{\lambda}, \quad \rho^2(\theta) = \frac{\lambda}{1 + \lambda}.$$

Then if $\hat{r}_i = 1 - \hat{p}_i$, $i = 1, 2$, putting $w = e_q$ in (5.1),

$$\hat{\mu}_c^e = \hat{p} + \left(\frac{\hat{p}_1\hat{r}_1}{\lambda n}\right)^{1/2} e_q\left(\frac{(\lambda n)^{1/2}(\hat{p}_1 - \hat{p}_2)}{(1 + \lambda)(\hat{p}_1\hat{r}_1)^{1/2}}\right)$$

or

$$\begin{aligned} \hat{\mu}_c^e &= \hat{p} \text{ if } |(\lambda n)^{1/2}(\hat{p}_1 - \hat{p}_2)/(\hat{p}_1\hat{r}_1)^{1/2}(1 + \lambda)| \le q \\ &= \hat{p}_1 - q \operatorname{sgn}(\hat{p}_1 - \hat{p}_2)(\lambda n)^{-1/2}(\hat{p}_1\hat{r}_1)^{1/2} \text{ otherwise.} \end{aligned} \tag{5.3}$$

This yields, by (5.1), for quadratic loss, a relative loss in risk of

$$\sigma_1^{-2}(\theta) \sup_a M(\theta, a, w) - 1 = q^2/(1 + \lambda) \tag{5.4}$$

while the relative savings in risk are

$$1 - \sigma_1^{-2}(\theta) \sup_{V(\theta)} M(\theta, a, w) = (1 - m_0(e_q))/(1 + \lambda). \tag{5.5}$$

Clearly we can extend this approach to confidence intervals and the p-variate case. What we are doing should be clear from the examples. We essentially interpolate between the M.L.E.'s of $\mu(\theta)$ under $\mathscr{M}_0$ and $\mathscr{M}_1$ using weights which are functions of Wald's form of the test statistic for H: $\mu(\theta) \in \mu(\Theta_0)$ vs. K: $\mu(\theta) \in \mu(\Theta_1)$.

When we consider the limit of ordinary risks $M(\theta, \{\delta_n\})$ we find that procedures (5.1) generally exhibit a discontinuity at points of Θ_0, i.e. convergence of the risk is not uniform. This is reminiscent of Hodges' example of a super efficient estimate which is essentially a pretest estimate corresponding to a sequence of levels tending to 0. However the Hodges procedure has infinite relative loss in risk whereas we propose to pay a small price in the relative loss in exchange for improved behaviour on Θ_0.

6. Conclusions: Open questions.

(1) We have applied robustness ideas to derive what we judge are useful biased estimates in the estimation of single parameters under a simple model $\mathscr{M}_0$ when we want to guard against deviations towards a larger model $\mathscr{M}_1$. The solutions involve both an approximation to the optimality principle and in general a large sample approximation. Tables 1 and 2 show that the first approximation is not serious for quadratic loss and the solutions give reasonable confidence intervals. The adequacy of the large sample approximation remains to be assessed in different models by obtaining approximate solutions of the Berger-Bickel type to the exact model, where possible.

(2) In the p-variate case, even approximate solutions can only be calculated in special cases and their structure depends on the loss function. It may be appropriate to apply Steinian "pulling in" within the simple model towards a yet simpler model as well as further "pulling in" towards the simple model itself. Alternatively, if we do not believe that losses from errors made in estimation of different components of μ should be combined it may still make sense to apply pulling in towards $\mathscr{M}_0$ on each component individually.

(3) This approach is applicable, in principle, to large sample problems when $\mathscr{M}_1$ is nonparametric. For example, suppose we want to estimate features of distributions such as medians, means, or even the whole distribution or its density. Our approach suggests reasonable ways of interpolating between estimates based on parametric assumptions and nonparametric estimates.

(4) Typically we have more than one simple candidate model $\mathscr{M}_0$. It would be very interesting to obtain reasonable estimates of $\mu(\theta)$ which do well at each member of a set of simple models while still performing adequately at a super model $\mathscr{M}_1$.

(5) This work is closely connected with the recent studies of Marazzi (1980) and Berger (1982) on robust Bayesian inference. See also the thesis of Y. Ritov (1982) and Masreliez and Martin (1977). Problem (P) is precisely of that form, minimize the Bayes risk for a prior degenerate at $\{0\}$ subject to a bound on the maximum risk—interpreted as the worst that misspecification of the prior can do. On the other hand, if in our original problem we replace the maximum risk over $\mathscr{M}_0$ by an average, we are again in the robust Bayesian framework. We prefer not to try to specify prior distributions. Our point is just that a possibly naive belief in a simpler model can be catered to with reasonable safety.

Acknowledgement. I am grateful to B. Efron and P. Huber for helpful conversations and A. Marazzi for the calculation of the lower bounds.

REFERENCES

ANDERSON, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set. *Proc. Amer. Math. Soc.* **6** 170–176.

BANCROFT, T. A. and HAN, C. P. (1977). Inference based on conditional specification: a note and a bibliography. *Internat. Statist. Rev.* **45** 117–128.

BERGER, J. (1982). Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results. *Statistical Decision Theory and Related Topics III.* S. S. Gupta and J. Berger, eds. Academic, New York.

BICKEL, P. J. (1983). Minimax estimation of the mean of a normal distribution subject to doing well at a point. *Recent Advances in Statistics, Festschrift for H. Chernoff* 511–528. H. Rizvi and D. Siegmund, eds. Academic, New York.

BROWN, B. W. (1980). The crossover experiment for clinical trials. *Biometrics* **36** 69–80.

CHEN, S. Y. (1983). Restricted risk Bayes estimation for the mean of a multivariate normal distribution. Tech. Report 83-33, Purdue University.

HODGES, J. L. JR. and LEHMANN, E. L. (1952). The use of previous experience in reaching statistical decisions. *Ann. Math. Statist.* **23** 396–407.

HUBER, P. J. (1977). Robust covariances. *Statistical Decision Theory and Related Topics III.* S. S. Gupta and J. Berger, Eds. Academic, New York.

KIEFER, J. (1957). Invariance, minimax sequential estimation and continuous time processes. *Ann. Math. Statist.* **28** 573–601.

LECAM, L. (1969). Théorie asymptotique de la décision statistique. Presses de l'Université de Montréal.

MARAZZI, A. (1980). Robust Bayesian estimation for the linear model. Technical Report No. 27. E.T.H. Zurich.

MARAZZI, A. (1982). On constrained minimization of the Bayes risk for the linear model. Technical Report No. 34. E.T.H. Zurich.

MASRELIEZ, C. J. and MARTIN, R. D. (1977). Robust Bayesian estimation for the linear model and robustifying the Kalman filter. *I.E.E.E. Trans. Automat. Control* **AC-22** June 1977. 361–371.

MORRIS, C., RADHAKRISHNAN, R. and SCLOVE, S. L. (1972). Nonoptimality of preliminary test estimators for the mean of a multivariate normal distribution. *Ann. Math. Statist.* **43** 1481–1490.

MOSTELLER, F. (1948). On pooling data. *J. Amer. Statist. Assoc.* **43** 231–242.

RITOV, Y. (1982). Robust quasi Bayesian inference. Thesis, Hebrew University, Jerusalem.

SACKS, J. and YLVISAKER, D. (1978). Linear estimation for approximately linear models. *Ann. Statist.* **6** 1122–1137.

SCHEFFÉ, H. (1959). *The Analysis of Variance.* Wiley, New York.

SEN, P. K. (1979). Asymptotic properties of maximum likelihood estimators based on conditional specification. *Ann. Statist.* **7** 1019–1033.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

Chapter 3
Asymptotic Theory

Qi-Man Shao

3.1 Introduction to Four Papers on Asymptotic Theory

3.1.1 General Introduction

Asymptotic theory plays a fundamental role in the developments of modern statistics, especially in the theoretical analysis of new methodologies. Some asymptotic results may borrow directly from the limit theory in probability, but many need deep insights of statistical contents and more accurate approximations, which have in turn fostered further developments of limit theory in probability. Peter Bickel has made far-reaching and wide-ranging contributions to modern statistics. He is a giant in theoretical statistics. In asymptotic theory, besides his contributions to bootstrap and high-dimensional statistical inference, in this paper I shall focus on four of his seminal papers on asymptotic expansions and Bartlett correction for Bayes solutions, likelihood ratio statistics and maximum-likelihood estimator for general hidden Markov models. The papers will be reviewed in chronological order.

3.1.2 Asymptotic Theory of Bayes Solutions

The paper of Bickel and Yahav (1969) deals with the asymptotic theory of Bayes solutions in estimation and hypothesis testing. It proves that Bayes estimates arising from a loss function are asymptotically efficient and that the mean of the posterior distribution is asymptotically normal, which confirms a long time statistical folklore.

Q.-M. Shao (✉)
Department of Statistics, Chinese University of Hong Kong,
Shatin, Kowloon, Hong Kong
e-mail: maqmshao@ust.hk

J. Fan et al. (eds.), *Selected Works of Peter J. Bickel*, Selected Works in Probability and Statistics 13, DOI 10.1007/978-1-4614-5544-8_3,

The results also significantly extend some early work of Le Cam. More importantly, the paper provides asymptotic expansions for the posterior risk in the estimation problem. The expansion can be viewed in the same sprit of Badahur's work, which is now commonly called *the Bahadur representation*. It is noted that Bickel and Yahav derived the expansion from an entirely different viewpoint.

The method, setup and results in Bickel and Yahav (1969) have a significant impact to the later wok in this area directly and indirectly. For example, Yuan (2009) proposed a joint estimation procedure in which some of the parameters are estimated Bayesian, and the rest by the maximum-likelihood estimator in the same parametric model. The proof of the consistency of the hybrid estimate is based on the method in Bickel and Yahav (1969). The paper of He and Shao (1996) on the Bahadur expansion for M-estimators follows a similar setup as Bickel and Yahav (1969). Belloni and Chernozhukov (2009) also follow the setup of B-Y and extend some of their results. The results of Bickel and Yahav (1969) have considerable applications in asymptotic sequential analysis. For recent results and extensions on this topic we refer to Hwang (1997), Ghosal (1999), and Belloni and Chernozhukov (2009) and references therein.

3.1.3 The Bartlett Correction

The Bartlett (1937) correction is a scalar transformation applied to the likelihood ratio (LR) statistic that yields a new improved test statistic which has a chi-squared null distribution to order $O(1/n)$. This represents a clear improvement of $O(1)$ for the the original LR statistic. A general frame work for Bartlett corrections was proposed by Lawley (1956). One can refer to Cribari-Neto and Cordeiro (1996) and Jensen (1993) for surveys on Bartlett corrections. The Bartlett correction is also closely related to Edgeworth expansions and saddlepoint approximations.

The main contributions of Bickel and Ghosh (1990) are twofolds: (1) it gives a generalization of Efron's (1985) result to vector parameters and applies this extension to establish the validity of Bartlett's correction to order $n^{-3/2}$, in particularly, verifies rigorously Lawley's (1956) result giving the order of the error in the Bartlett correction as $O(n^{-2})$; (2) it gives Bayesian analogues of both of above results that provide a key to understanding the Bartlett phenomenon.

The Bayesian idea in Bickel and Ghosh (1990) is creative. This enables to clear up mysteries such as why the Wald's or Rao's statistic is not Bartlett correctible and to explore the duality between the Bayesian and the frequentist setup. Let $X = (X_1, \ldots, X_n)$ be a vector of observations with joint density $p(x, \theta)$, $\theta = (\theta^1, \ldots, \theta^p) \in \Theta$ open in $\mathbf{R}^p$. For given θ, let $\hat{\theta}_0$ be the unrestricted MLE and $\hat{\theta}_j$ be the MLE of θ when $\theta^1, \ldots, \theta^j$ are fixed, i.e.,

$$l(\hat{\theta}_j) = \max\left\{l(\tau) : \tau^1 = \theta^1, \ldots, \tau^j = \theta^j\right\}, \ 1 \le j \le p.$$

Assume that these quantities exist and are unique and define $T = T(\theta, \mathbf{X}) = (T^1, \ldots, T^p)$ as the signed square roots of the likelihood ratio statistics, where

$$T_j = n^{1/2}\left\{2\left[l(\hat{\theta}_{j-1}) - l(\hat{\theta}_j)\right]\right\}^{1/2}\left(\hat{\theta}^j_{j-1} - \theta^j\right).$$

The Bayesian route begins with putting a prior density π on Θ. Let P denote the joint distribution of $(\theta, \mathbf{X})$ and $P(\cdot|\mathbf{X})$ the conditional (posterior) probability distribution of $(\theta, \mathbf{X})$ given $\mathbf{X}$. The posterior density of $\sqrt{n}(\theta - \hat{\theta})$ is given by

$$\pi(h|\mathbf{x}) \equiv \exp\left\{l(\hat{\theta} + rh) - l(\hat{\theta})\right\}\pi(\hat{\theta} + rh)/N(\mathbf{X}),$$

where $N(\mathbf{X}) = \int \exp\{l(\hat{\theta} + rh) - l(\hat{\theta})\}\pi(\hat{\theta} + rh)dh$. Let $\pi_T(t|\mathbf{X})$ denote the posterior density of T and $\phi(t) = (2\pi)^{-p/2}\exp\{\sum_{i=1}^p (t_i)^2/2\}$ be the standard p variate normal density. Bickel and Ghosh's (1990) first result is that, under certain "Bayesian" regularity conditions,

$$E_P \int |\pi_T(t|\mathbf{X}) - \pi_2(t, \mathbf{X})|dt = O(n^{-3/2}),$$

where

$$\pi_2(t, \mathbf{X}) = \phi(t)\left\{1 + P_{21}(\mathbf{X}, \pi)n^{-1/2} + P_{22}(\mathbf{X}, \pi)n^{-1} + Q_2(n^{-1/2}t)\right\}I\{\mathbf{X} \in S\}$$

for $t \in \mathbb{R}^p$. Here Q_2 is a polynomial in $n^{-1/2}t$ of degree 2 without a constant term and S is a set such that $P(\mathbf{X} \notin S) = O(n^{-3/2})$.

The second result in Bickel and Ghosh (1990) is to use above expansion in the Bayesian setup to establish the corresponding result in the frequentist case. Under certain frequentist conditions in an analogous fashion, the characteristic function of the density of T, $p_T(t|\theta)$, differs from that of $\mathcal{N}(n^{-1/2}R_{1j}, I_p + n^{-1}(2R_{2ij} - R_{i1}R_{1j}))$ by $O(n^{-3/2})$.

In addition, an asymptotic expansion for the distribution of the p deviances statistics up to $O(n^{-2})$ is also derived. More specifically, the vectors of deviances $D = (D^1, \ldots, D^p)$ and its Bartlett corrected version $\tilde{D} = (\tilde{D}^1, \ldots, \tilde{D}^p)$ are given by

$$D^j = (T^j)^2 = 2n\left[l(\hat{\theta}_{j-1}) - l(\hat{\theta}_j)\right]$$

and

$$\tilde{D}^j = D^j/(1 + 2n^{-1}Q_{2jj}).$$

Then, under regularity conditions, with error $O(n^{-1})$, the joint distribution of D is that of p independent χ^2_1, while for $\tilde{D}$ the same claims holds with error $O(n^{-2})$.

Note that the required assumptions, i.e., the regularity conditions in both Bayesian and frequentist settings, might appear rather strong. However, by examining several cases, including independent non-identically distributed and Markov dependent observations in Bickel et al. (1985) and exponential families in some regression and GLIM models, they hold quite generally. Indeed, similar type of regularity conditions were also assumed or served as basic assumptions in different problems for the validity of Edgeworth expansions. See, for example, Datta et al. (2000), Mukerjee and Reid (2001), Fang and Mukerjee (2006) and Fraser and Rousseau (2008).

Bickel and Ghosh (1990) addressed the rationale that why such accurate expansions, which hold for the likelihood ratio, would fail for Wald's or Rao's statistic. Moreover, a nice feature of their approach is that calculations are kept to a minimum such that the phenomena are transparent. Bickel and Ghosh's Bayesian results may be viewed as technical lemmas for proving frequentist theorems. This ingenious Bayesian argument has come a widely used statistical methodology after the appearance of Bickel and Ghosh's paper. In particular, Fan et al. (2000) applied a similar argument to provide a geometric understanding of the classical Wilks theorem as well as a useful extension of the likelihood ratio theory. For further extensions and related work, we refer to Dudley and Haughton (2002) and Schennach (2007).

It is noted that the Bartlett correction provides a measure of absolute error for the approximation. Since the tail probability of chi-squared distribution is exponentially decay, it would be interesting to see if a similar result holds for the relative error, or if a Cramér type moderate deviation with error $O(n^{-2})$ is valid.

3.1.4 Asymptotic Distribution of the Likelihood Ratio Statistic in Mixture Model

Mixture models are useful in describing data from a population that is suspected to be composed of a number of homogeneous subpopulations. The models have been used in econometrics, biology, genetics, medicine, agriculture, zoology, and population studies.

Bickel and Chernoff (1993) is the first paper that gives the asymptotic distribution of the likelihood ratio statistic in normal mixture model. Let $X_1, X_2, \cdots, X_n$ be i.i.d. $N(0,1)$ random variables and set $M_n^* = \sup_t S_n^*(t)$, where

$$S_n^*(t) = n^{-1/2} \sum_{i=1}^{n} y^*(X_i, t),$$

$$y^*(x,t) = (e^{tx-t^2/2} - 1 - tx)/(e^{t^2} - 1 - t^2)^{1/2}.$$

Hartigan (1984) proved that M_n^{*2} is stochastically equal to the logarithm of the likelihood ratio test statistic based on a normal mixture model $(1-p)N(0,1) + pN(\theta, 1)$ and that $M_n^* \to \infty$ in probability. Hartigan also conjectured that $M_n^* = O((\log_2 n)^{1/2})$, where $\log_2 n = \log(\log n)$. In Bickel and Chernoff (1993), Bickel and Chenoff confirm the Hartigan conjecture and more importantly, give an explicit asymptotic distribution of M_n^* as $n \to \infty$

$$P(V_n \le v) \to \exp(-e^{-v}), \tag{3.1}$$

where

$$V_n = M_n^*(\log_2 n)^{1/2} - \log_2 n + \log(\sqrt{2}\pi).$$

The main idea of the proof of (3.1) is to first deal with a simpler process

$$S_n(t) = n^{-1/2} \sum_{i=1}^{n} \left(e^{tX-i-t^2/2} - 1 \right) e^{-t^2/2},$$

which can be approximated by a Gaussian process by using the strong approximation (see Komlos et al. 1975; Csörgő and Révész 1981), and then apply the asymptotic distribution for maximal of stationary Gaussian process (Leadbetter et al. 1983). The approach in Bickel and Chernoff (1993) is applicable to other mixture and change point problems as the strong approximation works not only for sums of independent random variables but also for a lot of dependent variables. The paper has also inspired many follow-up studies on this topic, including Chen et al. (2004) and Charnigo and Sun (2004) and many others.

It is noted that the limiting distribution in (3.1) is called the extreme distribution of type I. It is commonly believed that the rate of convergence is extremely slow. Liu et al. (2008) show that an "intermediate approximation" may give a much faster rate of convergence. We also remark that a useful approach to deal with the asymptotic distribution of extreme values is Stein-Chen method, see Arratia et al. (1989).

3.1.5 Hidden Markov Models

A hidden Markov model (HMM) is a discrete-time stochastic process $\{(X_k, Y_k)\}$ such that (1) $\{X_k\}$ is a finite-state Markov chain, and (2) given $\{X_k\}$, $\{Y_k\}$ is a sequence of conditionally independent random variables. Hidden Markov models have been successfully applied in various areas of dependent data analysis, including speech recognition (Rabiner 1989), neurophysiology (Fredkin and Rice 1992), biology (Leroux and Puterman 1992; Holzmann et al. 2006), econometrics (Rydén et al. 1998) and medical statistics (Albert 1991) or biological sequence alignment (Arribas-Gil et al. 2006).

Inference for HMMs was initiated by Baum and Petrie (1966) for the case when $\{Y_k\}$ takes values in a finite set, where consistency and asymptotic normality of the maximum-likelihood estimator (MLE) are proved. For general HMMs, Lindgren (1978) constructed consistent and asymptotically normal estimators of the parameters determining the conditional densities of Y_n given X_n. Leroux (1992) proved consistency of the MLE for general HMMs under mild conditions, and Bickel and Ritov (1996) proved the local asymptotic normality, by using a quite long tedious analysis with more than 20 lemmas. Bickel et al. (1998) is the first article to establish rigorously the asymptotic normality of the MLE for general HMMs, which, together with the consistency proved by Leroux (1992), provides theoretical foundation for the validity and effectiveness of MLE. The impact of their paper is substantial. The results are obtained under mild regularity conditions of Cramér type that could not be weaken markedly and serve as basic assumptions in most subsequent statistical methodologies related to asymptotic studies for HMMs. Their results also provide possibilities on inference for a great many HMM related statistical problems due to the intrinsic nature of MLE.

Let $\{X_k, k \geq 1\}$ be a stationary Markov chain on $\{1,\ldots,K\}$ with transition probabilities $\alpha_\vartheta(a,b)$, where the parameter $\vartheta \in \Theta \subseteq \mathbb{R}^q$. Also let $\{Y_k\}$ be an $\mathscr{Y}$-valued sequence such that given $\{X_k\}$, $\{Y_k\}$ is a sequence of conditionally independent random variables with Y_n having conditional density $g_\vartheta(t|X_n)$. The MLE, denoted by $\hat{\vartheta}_n$, maximizes the joint density of $(Y_1,\ldots,Y_n)$, say $p_\vartheta(y_1,\ldots,y_n)$, over the parameter set Θ. The true parameter is denoted by ϑ_0. Bickel et al. (1998) showed that under Cramér-type conditions at ϑ_0 and ergodicity of $\{\alpha_{\vartheta_0}(a,b)\}$,

$$n^{1/2}\big(\hat{\vartheta}_n - \vartheta_0\big) \to \mathscr{N}(0, \mathscr{I}_0^{-1}), \;\; P_{\vartheta_0}\text{-weakly as } n \to \infty,$$

where $\mathscr{I}_0$ denotes the Fisher information matrix for $\{Y_k\}$ and is nonsingular.

In order to establish above main result, they first proved a central limit theorem for the score function (i.e. $L_n(\vartheta) = \log p_\vartheta(Y_1,\ldots,Y_n)$) at ϑ_0 with limit covariance matrix $\mathscr{I}_0$, that is,

$$n^{-1/2} D L_n(\vartheta_0) \to \mathscr{N}(0, \mathscr{I}_0^{-1}), \;\; P_{\vartheta_0}\text{-weakly as } n \to \infty.$$

A second result was a uniform law of large numbers for the Hessian of the log-likelihood, i.e.

$$n^{-1} D^2 L_n(\hat{\vartheta}_n) \to -\mathscr{I}_0 \;\text{ in } P_{\vartheta_0}\text{-probability}$$

as $n \to \infty$. Here D and D^2 form the gradient and the Hessian, respectively.

The paper of Bickel et al. (1998) furnishes the mathematical tools to studying HMMs and also opens a door for developing asymptotic theory of other statistical objects based on HMMs. For instance, Bickel et al. (2002) gave explicit expressions for derivatives and expectations of the log-likelihood function of HMMs and obtain second order asymptotic normality. Douc and Matias (2001) considered the consistency and asymptotic normality of the MLE for a possibly non-stationary hidden Markov model. After a relatively mature development on the statistical inference, Fuh (2004) studied the issue of hypothesis testing for HMM, in particular the problem of sequential probability ratio tests for parametrized HMMs. More recently, Dannemann and Holzmann (2009) discussed how the relevant asymptotic distribution theory for the likelihood ratio test when the true parameter is on the boundary can be extended from the i.i.d. situation to HMMs. Bickel et al. (1998) has inspired many subsequent work, including Douc et al. (2004), Vandekerkhove (2005), Fuh and Hu (2007), Anderson and Rydén (2009) and Sun and Cai (2009), among others. One can refer to Moulines et al. (2005) for recent developments in this area.

Acknowledgements Partially supported by Hong Kong RGC – CRG 603710.

References

Anderson S, Rydén T (2009) Subspace estimation and prediction methods for hidden Markov models. Ann Stat 37:4131–4152

Arratia R, Goldstein L, Gordon L (1989) Two moments suffice for Poisson approximation: the Chen-Stein method. Ann Probab 17:9–25

Arribas-Gil A, Gassiat E, Matias C (2006) Parameter estimation in pair-hidden Markov models. Scand J Stat 33(4):651–671

Albert PS (1991) A two-state Markov mixture model for a time series of epileptic seizure counts. Biometrics 47:1371–1381

Baum LE, Petrie T (1966) Statistical inference for probabilistic functions of finite state Markov chains. Ann Math Stat 37:1554–1563

Bartlett MS (1937) Properties of sufficiency and statistical tests. Proc R Soc Lond Ser A 160:268–282

Belloni A, Chernozhukov V (2009) On the computational complexity of MCMC-based estimators in large samples. Ann Stat 37:2011–2055

Bickel PJ, Chernoff H (1993) Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In: Ghosh JK, Mitra SK, Parthasarthy KR, Prakasa Rao BLS (eds) Statistics and probability: a Raghu Raj Bahadur festschrift. Wiley Eastern Limited, New Delhi, pp 83–96

Bickel PJ, Ghosh JK (1990) A decomposition for the likelihood ratio statisti and the Bartlett correction – a Bayesian argument. Ann Stat 18:1070–1090

Bickel PJ, Ritov Y (1996) Inference in hidden Markov models I: local asymptotic normality in the stationary case. Bernoulli 2:199–228

Bickel PJ, Yahav JA (1969) Some contribuions to the aymptotic theory of Bayes solutions. Z Wahrsch verw Geb 11:257–276

Bickel PJ, Götze F, van Zwet WR (1985) A simple analysis of third-order efficiency of estimates. In: Le Cam L, Olshen RA (eds) Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, vol 2. Wadsworth, Belmont, pp 749–768

Bickel PJ, Ritov Y, Rydén T (1998) Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. Ann Stat 26:1614–1635

Bickel PJ, Ritov Y, Rydén T (2002) Hidden Markov model likelihoods and their derivatives behave like i.i.d. ones. Ann Inst H Poincaré Probab Stat 38:825–846

Charnigo R, Sun J (2004) Testing homogeneity in a mixture distribution via the L^2 distance between competing models. J Am Stat Assoc 99:488–498

Chen H, Chen J, Kalbfleisch JD (2004) Testing for a finite mixture model with two components. J R Stat Soc B 66:95–115

Cribari-Neto F, Cordeiro GM (1996) On Bartlett and Bartlett-type corrections. Econ Rev 15:339–367

Csörgő M, Révész P (1981) Strong approximations in probability and statistics. Academic, New York

Dannemann J, Holzmann H (2008) Likelihood ratio testing for hidden Markov models under non-standard conditions. Scand J Stat 35:309–321

Datta GS, Mukerjee R, Ghosh M, Sweeting TJ (2000) Bayesian prediction with approximate frequentist validity. Ann Stat 28:1414–1426

Douc R, Matias C (2001) Asymptotics of the maximum likelihood estimator for general hidden Markov models. Bernoulli 7:381–420

Douc R, Moulines É, Rydén T (2004) Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. Ann Stat 32:2254–2304

Dudley RM, Haughton D (2002) Asymptotic normality with small relative errors of posterior probabilities of half-spaces. Ann Stat 20:1311–1344

Fan J, Hung H-N, Wong W-H (2000) Geometric understanding of likelihood ratio statistics. J Am Stat Assoc 95:836–841

Fang K-T, Mukerjee R (2006) Empirical-type likelihoods allowing posterior credible sets with frequentist validity: higher-order asymptotics. Biometrika 93:723–733

Fraser DAS, Rousseau J (2008) Studentization and deriving accurate p-values. Biometrika 95:1–16

Fredkin DR, Rice JA (2001) Fast evaluation of the likelihood of an HMM: ion channel currents with filtering and colored noise. IEEE Transactions on Signal Processing 49, p. 625

Fuh C (2004) On Bahadur efficiency of the maximum likelihood estimator in hidden Markov models. Stat Sin 14:127–154

Fuh C, Hu I (2007) Estimation in hidden Markov models via efficient importance sampling. Bernoulli 13:492–513

Ghosal S (1999) Asympotic normality of posterior distributions in high-dimensional linear models. Bernoulli 5:315–331

Hartigan JA (1985) A failure of likelihood ratio asymptotics for normal mixtures. In: Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer. Wadsworth Advanced Books, Monterey; Institute of Mathematical Statistics, Hayward

He XM, Shao QM (1996) A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. Ann Stat 24:2608–2630

Holzmann H, Munk A, Suster M, Zucchini W (2006) Hidden Markov models for circular and linear-circular time series. Environ Ecol Stat 13:325–347

Hwang LC (1997) A bayesian approach to sequential estimation without using the prior information. Seq Anal 16:319–343

Jensen JL (1993) A historical sketch and some new results on the improved likelihood statistic. Scand J Stat 20:1–15

Komlos J, Major P, Tusnady G (1975) An approximation of partial sums of independent r.v.s. and the sample distribution function. Z Wahrsch verw Geb 32:111–131

Lawley DN (1956) A general method for approximating to the distribution of the likelihood ratio criteria. Biometrika 43:295–303

Leadbetter MR, Lindgren G, Rootzen H (1983) Extremes and related properties of random sequences. Springer, New York

Leroux BG (1992) Maximum-likelihood estimation for hidden Markov models. Stoch Process Appl 40:127–143

Leroux BG, Puterman ML (1992) Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. Biometrics 48:545–548

Liu WD, Lin ZY, Shao QM (2008) The asymptotic distribution and Berry-Esseen bound of a new test for independence in high dimension with an application to stochastic optimization. Ann Appl Probab 18:2337–2366

Moulines E, Cappé O, Rydén T (2005) Inference in hidden Markov models. Spinger, New York/London

Mukerjee R, Reid N (2001) Second-order probability matching priors for a parametric function with application to Bayesian tolerance limits. Biometrika 88:587–592

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77:257–286

Rydén T, Terävirta T, Åsbrink S (1998) Stylized facts of daily return series and the hidden Markov model. Journal of Applied Econometrics, 13(3):217–244

Schennach SM (2007) Point estimation with exponentially tilted empirical likelihood. Ann Stat 35:634–672

Sun W, Cai T (2009) Large-scale multiple testing under dependence. J R Stat Soc B 71:393–424

Vandekerkhove P (2005) Consistent and asymptotically normal parameter estimates for hidden Markov mixtures of Markov models. Bernoulli 11:103–129

Yuan A (2009) Bayesian frequentist hybrid inference. Ann Stat 37:2458–2501

Z. Wahrscheinlichkeitstheorie verw. Geb. 11, 257—276 (1969)

Some Contributions to the Asymptotic Theory of Bayes Solutions

P. J. BICKEL* and J. A. YAHAV**

Received December 6, 1967

Summary. This paper deals with the asymptotic theory of Bayes solutions in (i) Estimation (ii) Testing when hypothesis and alternative are separated at least by an indifference region, under the assumption that the observations are independent and indentically distributed. The estimation results which are partial generalizations of results of LeCam begin with a proof of the convergence of the normalized posterior density to the appropriate normal density in a strong sense. From this result we derive the asymptotic efficiency of Bayes estimates obtained from smooth loss functions and in particular of the posterior mean. The last two theorems of this section deal with asymptotic expansions for the posterior risk in such estimation problems. The section on testing contains a limit theorem for the n-th root of the posterior risk under weak conditions on the prior and the loss function. Finally we discuss generalizations and some open problems.

1. Introduction

Our main purpose in writing this paper is to establish generalizations and give some simple extensions of the results on the asymptotic behavior of the Bayes posterior risk announced and proved in [3].

Section 2 deals with estimation of a real parameter in the presence of nuisance parameters, or more generally estimation of a real function g of a vector parameter $\boldsymbol{\theta}$. We assume we are given a Bayes prior density on the parameter space (including nuisance parameters). In general we take a decision theoretic point of view and suppose we are given a loss function $l(\boldsymbol{\theta}, d) = \tilde{l}(|g(\boldsymbol{\theta}) - d|)$ where the decision d is permitted to range over the real line. Our approach here as in our previous paper is basically that developed by LE CAM [12], [13] and [14] and independently by WOLFOWITZ [17]. Under our supplementary conditions we have been able to extend LE CAM's work to prove that Bayes estimates arising from loss functions l as above, where $l(t)$ behaves like a power of t near the origin and at infinity, are asymptotically efficient in the sense of CRAMÈR [5]. For instance under some conditions on g, in theorem 2.3, we show that the mean of the posterior distribution of $g(\boldsymbol{\theta})$ is asymptotically normally distributed about $g(\boldsymbol{\theta})$ with the appropriate variance. This result long a part of the statistical folklore does not seem yet to have appeared in the literature except as an abstract [6]. The main theorem of this section, 2.4, shows under various regularity conditions

---

* Part of this research was done while P. J. BICKEL was on leave at Imperial College, London. — This research was partially supported by National Science Foundation Grant GP-5059.

** This research was partially supported by National Science Foundation Grant GP-5705. Part of this research was done while J. A. YAHAV was visiting the department of Statistics at Stanford University.

that the Bayes posterior risk behaves like $n^{-\beta}$ where β, not surprisingly, depends on the behavior of $\tilde{l}$ near the origin. We also give some expansions for the posterior risk in Theorem 2.5. These limiting results have considerable application in asymptotic sequential analysis (c.f. [3], [4], and [22]). In section 3 we deal with the behavior of the Bayes posterior risk in testing disjoint hypotheses or hypotheses which are separated by an indifference region in which the losses due to taking the wrong decision are 0. In this case the posterior risk goes to 0 exponentially and its n-th root is related to the Kullback-Leibler information numbers. Again these results find application in asymptotic sequential analysis, and complement those of KIEFER and SACKS [11]. Though proceeding from an entirely different viewpoint our work in this section is also closely related to that of BAHADUR [1] and BAHADUR and BICKEL [2]. Section 4 contains a discussion of possible generalizations.

2. Estimation

Let $z_1, \ldots, z_n, \ldots$ be a sequence of independent identically distributed random variables (observations) defined on a measurable space $(\Omega, \mathfrak{A})$. Suppose that z_1 is distributed according to one of the probability laws $P_{\boldsymbol{\theta}}$ on $(R, \mathfrak{B})$ where R is the real line, $\mathfrak{B}$ is the Borel field. $\boldsymbol{\theta}$ ranges over an index set (parameter space) Θ. Throughout this paper we shall suppose that $P_{\boldsymbol{\theta}}$ is dominated by some σ finite measure μ on $(\Omega, \mathfrak{A})$ for all $\boldsymbol{\theta}$ and we denote the density of z_1 if $\boldsymbol{\theta}$ is true, $dP_{\boldsymbol{\theta}}/d\mu$, by $f(z, \boldsymbol{\theta})$. In addition, we denote $\log f(z, \boldsymbol{\theta})$ by $\Phi(z, \boldsymbol{\theta})$. As usual, $\log 0 = -\infty$. For simplicity we also refer to $P_{\boldsymbol{\theta}}$ when we wish to speak about the probability induced by the $\{z_i\}$ on $(R^\infty, \mathfrak{B}^\infty)$ the infinite dimensional product space.

Of course, we take $P_{\boldsymbol{\theta}_1} \neq P_{\boldsymbol{\theta}_2}$ if $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$. We suppose,

A 2.1. Θ may be identified with an open subset of R^k. This assumption is also used in section 3 but as will clearly be seen is given there mainly as a convenience. In fact, the argument of section 3 continues to be valid for any locally compact metric space, and as is indicated in that section can be further generalized. On the other hand A 2.1 is crucial for the present section. We suppose we are given a Bayes prior measure Ψ on Θ endowed with the Borel σ-field. We assume,

A 2.2. Ψ has a density ψ with respect to k dimensional Lebesgue measure. Moreover ψ is continuous, positive and bounded on Θ. We are interested in estimating a univariate function g of our vector parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. The structure we require on g is embodied in,

A 2.3. Let,

$$\operatorname{grad} g(\boldsymbol{\theta}) = \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_k}\right) \tag{2.1}$$

exist, be continuous, and bounded on Θ viz.

$$\sup\left\{\sum_{i=1}^{k}\left|\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_i}\right| : \boldsymbol{\theta} \in \Theta\right\} < \infty . \tag{2.2}$$

Moreover suppose $\operatorname{grad} g(\boldsymbol{\theta}) \neq \mathbf{0}$.

For theorem 2.5, in order to avoid unduly messy expansions we make the very strong assumption.

A 2.3′. The function g to be estimated is of the form

$$g(\boldsymbol{\theta}) = \sum_{i=1}^{k} a_i \theta_i \quad \text{for some non zero vector} \quad (a_1, \ldots, a_k).$$

To specify that this is an estimation situation we require suitable loss functions and we suppose we are given a function $\tilde{l}(t)$ on $[0, \infty)$ such that,

A 2.4. a) $\tilde{l} \geqq 0$

b) $\tilde{l}(0) = 0$

c) $\tilde{l}$ has a derivative $\tilde{l}'$ on $(0, \infty)$ which is positive. Moreover, $\tilde{l}'$ is continuous and bounded in a neighborhood of 0.

d) There exists $s \geqq 0$, $\gamma > 0$ such that

$$t^{-s}\tilde{l}'(t) \to \gamma \quad \text{if} \quad t \to 0. \tag{2.3}$$

e) There exists $1 \leqq r < \infty$, such that

$$\limsup_{t\to\infty} [\tilde{l}'(t)]\, t^{-(r-1)} < \infty. \tag{2.4}$$

For theorem 2.5 we will need the stronger,

A 2.4′. $$\tilde{l}(t) = \frac{\gamma t^{s+1}}{(s+1)}, \quad \text{where} \quad \gamma > 0, \quad 1 \leqq s < \infty.$$

Since our decision space $D = R$ we take our loss function $l(\boldsymbol{\theta}, d)$ to be given by,

$$l(\boldsymbol{\theta}, d) = \tilde{l}(|g(\boldsymbol{\theta}) - d|). \tag{2.5}$$

Our next assumption puts a further restriction on ψ in view of A 2.4. Let $\|\cdot\|$ be the usual Euclidean norm. Then,

A 2.5.

$$\int_{\Theta} \|\boldsymbol{\theta}\|^r \prod_{i=1}^{n} f(z_i, \boldsymbol{\theta})\, \psi(\boldsymbol{\theta})\, d\boldsymbol{\theta} < \infty \tag{2.6}$$

a.s. $P_{\boldsymbol{\theta}}$ for all $\boldsymbol{\theta}$. (In theorems 2.3—2.5 the "r" of A 2.4 and A 2.5 is the same. Assumption A 2.4 is irrelevant to theorem 2.2.)

This is a consequence of

$$\int_{\Theta} \|\boldsymbol{\theta}\|^r \psi(\boldsymbol{\theta})\, d\boldsymbol{\theta} < \infty. \tag{2.7}$$

The expression in (2.6) is > 0 a.s. $P_{\boldsymbol{\theta}}$ for all $\boldsymbol{\theta}$ in view of the positivity of ψ and the continuity of $f_{\boldsymbol{\theta}}$.

Let

$$\psi(\boldsymbol{\theta} \,|\, z_1, \ldots, z_n) = \psi(\boldsymbol{\theta}) \prod_{i=1}^{n} f(z_i, \boldsymbol{\theta}) \left[\int_{\Theta} \prod_{i=1}^{n} f(z_i, \mathbf{s})\, \psi(\mathbf{s})\, d\mathbf{s} \right]^{-1} \tag{2.8}$$

the posterior density of $\boldsymbol{\theta}$ given $z_1, \ldots, z_n$. The posterior density is clearly well defined and finite a.s. $P_{\boldsymbol{\theta}}$.

We define

(2.8) (a) $$Y_n = \min \left\{ \int_{\Theta} l(\boldsymbol{\theta}, d)\, \psi(\boldsymbol{\theta} \,|\, z_1, \ldots, z_n)\, d\boldsymbol{\theta} \colon\ d \in R \right\}$$

18*

where we suppose for each $z_1, \ldots, z_n$ the minimum is achieved and a measurable version $\hat{g}_n(z_1, \ldots, z_n)$ of the minimizing decision exists. Then Y_n is measurable and $\hat{g}_n$ satisfies,

$$\int_{[g(\boldsymbol{\theta})<\hat{g}_n]} \tilde{l}'(|\hat{g}_n - g(\boldsymbol{\theta})|)\, \psi(\boldsymbol{\theta}\,|\,z_1, \ldots, z_n)\, d\boldsymbol{\theta} = \int_{[g(\boldsymbol{\theta})>\hat{g}_n]} \tilde{l}'(|\hat{g}_n - g(\boldsymbol{\theta})|)\, \psi(\boldsymbol{\theta}\,|\,z_1, \ldots, z_n)\, d\boldsymbol{\theta}. \tag{2.9}$$

The validity of (2.9) follows from A 2.3, A 2.4, A 2.5 and the dominated convergence theorem by standard arguments.

This characterization of Bayes estimates was suggested to us by the work of Farrell [9].

Our final set of assumptions is classical.

A 2.6.

$$\frac{\partial \Phi(z_1, \boldsymbol{\theta})}{\partial \theta_i} \quad \text{and} \quad \frac{\partial^2 \Phi(z_1, \boldsymbol{\theta})}{\partial \theta_i\, \partial \theta_j}$$

exist and are continuous in $\boldsymbol{\theta}$ for almost all z.

A 2.7.

$$E_{\boldsymbol{\theta}}\left(\sup\left\{\left|\frac{\partial^2 \Phi(z_1, \boldsymbol{s})}{\partial \theta_i\, \partial \theta_j}\right| : \|\boldsymbol{s} - \boldsymbol{\theta}\| < \varepsilon(\boldsymbol{\theta}),\ \boldsymbol{s} \in \Theta\right\}\right) < \infty \tag{2.10}$$

for some $\varepsilon(\boldsymbol{\theta})$, and all $i, j, \boldsymbol{\theta}$. $E_{\boldsymbol{\theta}}$ as usual denotes that computation is carried out when $\boldsymbol{\theta}$ is true.

Again for theorem 2.5 we need,

A 2.6′. $\dfrac{\partial^4 \Phi(z_1, \boldsymbol{\theta})}{\partial \theta_i\, \partial \theta_j\, \partial \theta_r\, \partial \theta_l}$ exist and are continuous in $\boldsymbol{\theta}$ for all i, j, r, l, and

A 2.7′. a) $E_{\boldsymbol{\theta}}\left(\sup\left\{\left|\dfrac{\partial^4 \Phi(z_1, \boldsymbol{s})}{\partial \theta_i\, \partial \theta_j\, \partial \theta_r\, \partial \theta_l}\right| : \|\boldsymbol{s} - \boldsymbol{\theta}\| < \varepsilon(\boldsymbol{\theta}),\ s \in \Theta\right\}\right) < \infty$

for some $\varepsilon(\boldsymbol{\theta})$ and all $i, j, r, l, \boldsymbol{\theta}$.

b) $E_{\boldsymbol{\theta}}\left[\dfrac{\partial^3 \Phi(z_1, \boldsymbol{\theta})}{\partial \theta_i\, \partial \theta_j\, \partial \theta_k}\right]^2 < \infty$ for all i, j, k.

A 2.6 and A 2.7 imply,

$$E_{\boldsymbol{\theta}}\left(\frac{\partial \Phi(z_1, \boldsymbol{\theta})}{\partial \theta_i}\right) = 0 \quad \text{for} \quad 1 \leqq i \leqq k, \tag{2.11}$$

and

$$A_{ij}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left(\frac{\partial^2 \Phi(z_1, \boldsymbol{\theta})}{\partial \theta_i\, \partial \theta_j}\right) = -E_{\theta}\left[\frac{\partial \Phi(z_1, \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \Phi(z_1, \boldsymbol{\theta})}{\partial \theta_j}\right]. \tag{2.12}$$

Let

$$A(\boldsymbol{\theta}) = \langle A_{ij}(\boldsymbol{\theta}) \rangle \tag{2.13}$$

where $\langle \cdot \rangle$ denotes a matrix.

We require,

A 2.8. $- A(\boldsymbol{\theta})$ is positive definite for all $\boldsymbol{\theta}$. The additional assumption we need to deal with Y_n is,

A 2.9.

$$E_{\boldsymbol\theta}[\sup\{[\Phi(z_1, \boldsymbol s) - \Phi(z_1, \boldsymbol\theta)]: \| \boldsymbol s - \boldsymbol\theta\| \geqq \varepsilon,\ \boldsymbol s \in \Theta\}] < 0, \quad \text{for all} \quad \boldsymbol\theta \in \Theta \quad \text{and} \quad \varepsilon > 0. \tag{2.14}$$

Fix $\boldsymbol\theta = \boldsymbol\theta_0$. We define, if U is a compact neighborhood of $\boldsymbol\theta_0$, $\hat{\boldsymbol\theta}_n(z_1, \ldots, z_n)$ to be a value of $\boldsymbol\theta$ such that,

$$\sum_{i=1}^{n} \Phi(z_i, \hat{\boldsymbol\theta}_n(z_1, \ldots, z_n)) = \max\left\{\sum_{i=1}^{n} \Phi(z_i, \boldsymbol\theta): \boldsymbol\theta \in U\right\}. \tag{2.15}$$

By lemma 3 of [14] we may choose $\hat{\boldsymbol\theta}_n$ measurable.

Of course, $\hat{\boldsymbol\theta}_n$ depends on U. We ask merely that U be such that the conclusion of the following lemma is satisfied. From this point on we shall for convenience stop indicating that statements hold only a.s. $P_{\boldsymbol\theta_0}$ when this is clear from the context. We have,

Lemma 2.1. *Suppose assumption* A 2.1, A 2.6 *and* A 2.7 *hold. Then, there exists a* $U(\boldsymbol\theta_0)$ *such that,*

$$\hat{\boldsymbol\theta}_n \to \boldsymbol\theta_0 \tag{2.16}$$

and there exists $N(z_1, \ldots, z_n, \ldots)$ *such that,*

$$\sum_{i=1}^{n} \operatorname{grad} \Phi(z_i, \hat{\boldsymbol\theta}_n) = 0 \tag{2.17}$$

for $n \geqq N$, *where*

$$\operatorname{grad} \Phi(z, \boldsymbol\theta) = \left(\frac{\partial \Phi(z, \boldsymbol\theta)}{\partial \theta_1}, \ldots, \frac{\partial \Phi(z, \boldsymbol\theta)}{\partial \theta_k}\right).$$

This lemma a generalization of lemma 3 of [14], is essentially contained in [13]; (see remark on p. 308). For convenience we sketch its proof.

Proof. Since $\boldsymbol\theta_0$ is in the interior of $U(\boldsymbol\theta_0)$, (2.17) will follow from (2.15), (2.16), and A 2.6.

Take $U(\boldsymbol\theta_0) = \{\boldsymbol\theta: \| \theta - \boldsymbol\theta_0\| \leqq \varepsilon/2\ (\boldsymbol\theta_0)\}$ where ε is given by A 2.7. By the multivariate Taylor theorem ([7] p. 186)

$$\begin{aligned}\Phi(z_1, \boldsymbol\theta) - \Phi(z_1, \boldsymbol\theta_0) &= \operatorname{grad} \Phi(z_1, \boldsymbol\theta_0) \cdot (\boldsymbol\theta - \boldsymbol\theta_0)' \\ &\quad + \tfrac{1}{2} \int_0^1 (\boldsymbol\theta - \boldsymbol\theta_0) \cdot A(z_1, \boldsymbol\theta_0 + \lambda(\boldsymbol\theta - \boldsymbol\theta_0)) \cdot (\boldsymbol\theta - \boldsymbol\theta_0)' \, d\lambda\end{aligned} \tag{2.18}$$

where the matrix A is given by,

$$A(z_1, \boldsymbol\theta) = \left\langle \frac{\partial \Phi(z_1, \boldsymbol\theta)}{\partial \theta_i \, \partial \theta_j} \right\rangle. \tag{2.19}$$

Applying A 2.6, A 2.7 and the dominated convergence theorem to (2.18) we see that for $\boldsymbol\theta \in U(\boldsymbol\theta_0)$

$$\begin{aligned}\lim_{\delta \to 0} E_{\boldsymbol\theta_0}[\sup\{\Phi(z_1, \boldsymbol\tau) - \Phi(z_1, \boldsymbol\theta_0): \| \boldsymbol\tau - \boldsymbol\theta\| < \delta,\ \boldsymbol\tau \in U(\boldsymbol\theta_0)\}] \\ = E_{\boldsymbol\theta_0}(\Phi(z_1, \boldsymbol\theta) - \Phi(z_1, \boldsymbol\theta_0)) < 0.\end{aligned} \tag{2.20}$$

If we now consider $U(\boldsymbol\theta_0)$ as our parameter space and restrict attention to $\boldsymbol\theta_0$,

A 2.9.

$$E_{\boldsymbol{\theta}}[\sup\{[\Phi(z_1, \boldsymbol{s}) - \Phi(z_1, \boldsymbol{\theta})]: \|\boldsymbol{s} - \boldsymbol{\theta}\| \geqq \varepsilon,\ \boldsymbol{s} \in \Theta\}] < 0, \quad \text{for all} \quad \boldsymbol{\theta} \in \Theta \quad \text{and} \quad \varepsilon > 0. \tag{2.14}$$

Fix $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. We define, if U is a compact neighborhood of $\boldsymbol{\theta}_0$, $\hat{\boldsymbol{\theta}}_n(z_1, \ldots, z_n)$ to be a value of $\boldsymbol{\theta}$ such that,

$$\sum_{i=1}^{n} \Phi(z_i, \hat{\boldsymbol{\theta}}_n(z_1, \ldots, z_n)) = \max\left\{\sum_{i=1}^{n} \Phi(z_i, \boldsymbol{\theta}) : \boldsymbol{\theta} \in U\right\}. \tag{2.15}$$

By lemma 3 of [14] we may choose $\hat{\boldsymbol{\theta}}_n$ measurable.

Of course, $\hat{\boldsymbol{\theta}}_n$ depends on U. We ask merely that U be such that the conclusion of the following lemma is satisfied. From this point on we shall for convenience stop indicating that statements hold only a.s. $P_{\boldsymbol{\theta}_0}$ when this is clear from the context. We have,

Lemma 2.1. *Suppose assumption* A 2.1, A 2.6 *and* A 2.7 *hold. Then, there exists a* $U(\boldsymbol{\theta}_0)$ *such that,*

$$\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0 \tag{2.16}$$

and there exists $N(z_1, \ldots, z_n, \ldots)$ *such that,*

$$\sum_{i=1}^{n} \operatorname{grad} \Phi(z_i, \hat{\boldsymbol{\theta}}_n) = 0 \tag{2.17}$$

for $n \geqq N$, *where*

$$\operatorname{grad} \Phi(z, \boldsymbol{\theta}) = \left(\frac{\partial \Phi(z, \boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial \Phi(z, \boldsymbol{\theta})}{\partial \theta_k}\right).$$

This lemma a generalization of lemma 3 of [14], is essentially contained in [13]; (see remark on p. 308). For convenience we sketch its proof.

Proof. Since $\boldsymbol{\theta}_0$ is in the interior of $U(\boldsymbol{\theta}_0)$, (2.17) will follow from (2.15), (2.16), and A 2.6.

Take $U(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta}: \|\theta - \boldsymbol{\theta}_0\| \leqq \varepsilon/2\ (\boldsymbol{\theta}_0)\}$ where ε is given by A 2.7. By the multivariate Taylor theorem ([7] p. 186)

$$\begin{aligned} \Phi(z_1, \boldsymbol{\theta}) - \Phi(z_1, \boldsymbol{\theta}_0) &= \operatorname{grad} \Phi(z_1, \boldsymbol{\theta}_0) \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \\ &\quad + \tfrac{1}{2} \int_0^1 (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \cdot A(z_1, \boldsymbol{\theta}_0 + \lambda(\boldsymbol{\theta} - \boldsymbol{\theta}_0)) \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\, d\lambda \end{aligned} \tag{2.18}$$

where the matrix A is given by,

$$A(z_1, \boldsymbol{\theta}) = \left\langle \frac{\partial \Phi(z_1, \boldsymbol{\theta})}{\partial \theta_i\, \partial \theta_j} \right\rangle. \tag{2.19}$$

Applying A 2.6, A 2.7 and the dominated convergence theorem to (2.18) we see that for $\boldsymbol{\theta} \in U(\boldsymbol{\theta}_0)$

$$\begin{aligned} \lim_{\delta \to 0} E_{\boldsymbol{\theta}_0}[\sup\{\Phi(z_1, \boldsymbol{\tau}) - \Phi(z_1, \boldsymbol{\theta}_0) : \|\boldsymbol{\tau} - \boldsymbol{\theta}\| < \delta,\ \boldsymbol{\tau} \in U(\boldsymbol{\theta}_0)\}] \\ = E_{\boldsymbol{\theta}_0}(\Phi(z_1, \boldsymbol{\theta}) - \Phi(z_1, \boldsymbol{\theta}_0)) < 0. \end{aligned} \tag{2.20}$$

If we now consider $U(\boldsymbol{\theta}_0)$ as our parameter space and restrict attention to $\boldsymbol{\theta}_0$, (2.16) follows by standard arguments on the consistency of maximum likelihood estimates in view of (2.20) and the compactness of $U(\boldsymbol{\theta}_0)$ (c.f. WALD [16]), or by the S.L.L.N. for Banach space valued random variables (c.f. [12]). ‖

Let,

$$\psi^*(\boldsymbol{t} \mid z_1, \ldots, z_n) = n^{-1/2} \psi(\boldsymbol{t} n^{-1/2} + \hat{\boldsymbol{\theta}}_n \mid z_1, \ldots, z_n) \tag{2.21}$$

the posterior density of $n^{1/2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)$.

LE CAM [14] has shown that $\psi^*(\boldsymbol{t} \mid z_1, \ldots, z_n)$ converges in the first mean to $\varphi(-A^{-1}(\boldsymbol{\theta}_0), \boldsymbol{t})$ where $\varphi(B, \boldsymbol{t})$ is the density of the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix B. The theorems of this section begin with a generalization of this result. Professor LE CAM has informed us that he has obtained a new proof of theorem 2.2 for the case he previously considered ($r = 0$) under much weaker assumptions than the ones we have stated.

Theorem 2.2. *Under assumptions* A 2.1—A 2.2, A 2.5—A 2.9 *if* $0 \leqq q \leqq r$.

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \|\boldsymbol{t}\|^q \left| \psi^*(\boldsymbol{t} \mid z_1, \ldots, z_n) - \varphi(-\mathrm{A}^{-1}(\boldsymbol{\theta}_0), \boldsymbol{t}) \right| d\boldsymbol{t} \to 0 . \tag{2.22}$$

From this will follow,

Theorem 2.3. *Under assumptions* A 2.1—A 2.9,

a) $$\hat{g}_n \to g(\boldsymbol{\theta}_0),$$

b) $$\mathfrak{L}_{\boldsymbol{\theta}_0}[n^{1/2}(\hat{g}_n - g(\boldsymbol{\theta}_0))] \to \mathfrak{N}(0, \sigma^2(g, \boldsymbol{\theta}_0))$$

where $\mathfrak{N}(0, \sigma^2)$ *is the normal distribution with mean* 0 *and variance* σ^2, *and,*

$$\sigma^2(g, \boldsymbol{\theta}) = -[\operatorname{grad} g(\boldsymbol{\theta})] \cdot A^{-1}(\boldsymbol{\theta}) \cdot [\operatorname{grad} g(\boldsymbol{\theta})]' \tag{2.23}$$

$\mathfrak{L}$. as usual stands for law.

Theorem 2.4. *Under assumptions* A 2.1—A 2.9 *if* $\beta = (s+1)/2$

$$n^{\beta} Y_n \to V(\boldsymbol{\theta}) \tag{2.24}$$

a.s. $P_{\boldsymbol{\theta}}$ *where,*

$$V(\boldsymbol{\theta}) = \gamma(s+1)^{-1} \sigma^{2\beta}(g, \boldsymbol{\theta}) \mu_{2\beta} \tag{2.25}$$

and μ_k *is the* k*-th absolute moment of the standard normal distribution.*

Our last theorem in this section gives an expansion for Y_n to one term behond $V(\boldsymbol{\theta}) n^{-\beta}$. Expansions to higher order terms are also possible but are rather complicated.

Theorem 2.5. *If assumptions* A 2.1—A 2.2, A 2.3′, A 2.4′, A 2.5, A 2.6′, A 2.7′ *hold, then,*

$$Y_n = n^{-\beta} \left\{ V(\boldsymbol{\theta}) + [2\pi \det A(\boldsymbol{\theta}_0)]^{-k/2} \left[\frac{S_{n1}(\boldsymbol{\theta})}{n} - V(\boldsymbol{\theta}) \frac{S_{n2}(\boldsymbol{\theta})}{n} \right] + 0(n^{-1/2}) \right\} \tag{2.26}$$

a.s. $P_{\boldsymbol{\theta}}$ *where*

$$S_{n1}(\boldsymbol{\theta}) = \frac{\gamma 1}{2(s+1)} \sum_{i=1}^{n} \int |g(\boldsymbol{t})|^{s+1} \{\boldsymbol{t}[A(z_i, \boldsymbol{\theta})] \boldsymbol{t}' + \boldsymbol{t} Q(z_i, \boldsymbol{\theta}) \boldsymbol{t}'\} \exp \frac{1}{2} \boldsymbol{t} A(\boldsymbol{\theta}) \boldsymbol{t}' \, d\boldsymbol{t}$$

where

$$Q(z_1, \boldsymbol{\theta}) = \left\| - E_{\boldsymbol{\theta}}\left(\operatorname{grad} \frac{\partial^2 \Phi(z_1, \boldsymbol{\theta})}{\partial\theta_r\, \partial\theta_s}\right) A^{-1}(\boldsymbol{\theta})\,[\operatorname{grad} \Phi(z_1, \boldsymbol{\theta})]' \right\|_{r,s}$$

and

$$S_{n2}(\boldsymbol{\theta}) = \frac{\Theta 1}{2} \sum_{i=1}^{n} \int \{\boldsymbol{t}[A(z_i, \boldsymbol{\theta}) - A(\boldsymbol{\theta})]\boldsymbol{t}' + \boldsymbol{t}\,Q(z_i, \boldsymbol{\theta})\,\boldsymbol{t}'\} \exp \frac{1}{2}\, \boldsymbol{t} A(\boldsymbol{\theta})\,\boldsymbol{t}'\, d\boldsymbol{t}.$$

Thus deviations of $n^\beta Y_n$ from $V(\boldsymbol{\theta})$ are of order $[n' \log\log n]^{-1/2}$. A generalization of this result to the case where $\tilde{l}$ admits a Taylor expansion around 0, $\tilde{l}(t) = \gamma_1 t + \gamma_2 t^2 + \cdots + \gamma_k t^k + 0(t^k)$ with $\min_j \gamma_j > 0$, and $\tilde{l}(t)$ satisfies A 2.5 b) may readily be formulated and proved. Similarly one may generalize to the case where $g(\boldsymbol{\theta})$ admits a Taylor expansion to p terms around $\boldsymbol{\theta} = \boldsymbol{0}$ with the remainder term uniformly of order $\|\boldsymbol{\theta}\|^p$. However, since a reparametrization of the parameter space to make $g(\boldsymbol{\theta})$ the first co-ordinate of our vector parameter is usually possible this generalization whose statement is extremely complicated hardly seems worthwhile.

A similar expansion for the loss incurred by using the Bayes estimate was obtained by Elfving in [21].

We proceed with the proof of theorem 2.2. We begin by defining

$$\nu_n(\boldsymbol{t}) = \exp \sum_{i=1}^{n} [\Phi(z_i, \boldsymbol{t} n^{-1/2} + \hat{\boldsymbol{\theta}}_n) - \Phi(z_i, \hat{\boldsymbol{\theta}}_n)]. \tag{2.27}$$

Then

$$\psi_n^*(\boldsymbol{t} \mid z_1, \ldots, z_n) = \nu_n(\boldsymbol{t})\, \psi(\boldsymbol{t} n^{-1/2} + \hat{\boldsymbol{\theta}}_n) \cdot \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \nu_n(\boldsymbol{t})\, \psi(\boldsymbol{t} n^{-1/2} + \hat{\boldsymbol{\theta}}_n)\, d\boldsymbol{t} \right]^{-1}. \tag{2.28}$$

As in [3] the theorem will follow if we can show,

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (1 + \|\boldsymbol{t}\|^r)\, \psi(\boldsymbol{t} n^{1/2} + \hat{\boldsymbol{\theta}}_n)\, | \nu_n(\boldsymbol{t}) - \varphi(-A^{-1}(\boldsymbol{\theta}_0), \boldsymbol{t})\, (2\pi)^{k/2} (\det[-A(\boldsymbol{\theta}_0)])^{-1/2} |\, d\boldsymbol{t} \to 0 \tag{2.29}$$

where det denotes determinant.

Since ψ is bounded, this would be a consequence of

$$\int \cdots \int_{\|\boldsymbol{t}\| \geq \delta^* n^{1/2}} \|\boldsymbol{t}\|^r\, \psi(\boldsymbol{t} n^{-1/2} + \hat{\boldsymbol{\theta}}_n)\, \nu_n(\boldsymbol{t})\, d\boldsymbol{t} \to 0 \tag{2.30}$$

for every $\delta^* > 0$, and

$$\int \cdots \int_{\|\boldsymbol{t}\| < \delta^* n^{1/2}} (1 + \|\boldsymbol{t}\|^r)\, H(\boldsymbol{t}, z_1, \ldots, z_n)\, d\boldsymbol{t} \to 0 \tag{2.31}$$

for some $\delta^* > 0$, where H is the expression within absolute value signs in (2.29).

We begin by proving,

Lemma 2.6. *Under assumptions* A 2.1—A 2.2 *and* A 2.5—A 2.9 *then* (2.30) *holds.*

Proof. If $\|\boldsymbol{t} n^{-1/2}\| \geq \delta^*$

$$n^{-1} \log \nu_n(\boldsymbol{t}) \leq n^{-1} \sum_{i=1}^{n} \sup \{\Phi(z_i, \boldsymbol{t}) - \Phi(z_i, \boldsymbol{\theta}_0) : \|\boldsymbol{t} - \boldsymbol{\theta}_0\| \geq \delta^* - \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|\} + n^{-1} \sum_{i=1}^{n} [\Phi(z_i, \boldsymbol{\theta}_0) - \Phi(z_i, \hat{\boldsymbol{\theta}}_n)]. \tag{2.32}$$

By lemma 2.1 and the S.L.L.N. (strong law of large numbers) the first term on the right hand side of (2.32) converges to

$$E_{\boldsymbol{\theta}_0}[[\sup\{\Phi(z_1, \boldsymbol{t}) - \Phi(z_1, \boldsymbol{\theta}_0) \colon \|\boldsymbol{t} - \boldsymbol{\theta}_0\| \geqq \delta^*\}] \leqq -\varepsilon(\delta^*) < 0.$$

The second term is by Taylor's theorem bounded in absolute value by,

$$n^{-1}\sum_{i=1}^{n} \sup\{\|\operatorname{grad}\Phi(z_i, \boldsymbol{t})\| \colon \|\boldsymbol{t} - \boldsymbol{\theta}_0\| \leqq \|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_n\|\} \cdot \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|. \tag{2.33}$$

Again applying A 2.6 and the S.L.L.N. we conclude that the first factor in (2.33) tends to $E_{\boldsymbol{\theta}_0}[(\|\operatorname{grad}\Phi(z_i, \boldsymbol{\theta}_0)\|)$, while the second factor tends to 0 by lemma 2.1.

We see that,

$$\sup\{\nu_n(\boldsymbol{t}) \colon \|\boldsymbol{t}\| \geqq \delta^* n^{1/2}\} \sim \exp - n\,\varepsilon(\delta^*). \tag{2.34}$$

Therefore,

$$\begin{aligned} &\int \cdots \int_{\|\boldsymbol{t}\| \geqq \delta^* n} \|\boldsymbol{t}\|^r \psi(\boldsymbol{t} n^{-1/2} + \hat{\boldsymbol{\theta}}_n)\, \nu_n(\boldsymbol{t})\, d\boldsymbol{t} \\ &\leqq n^{\frac{r+1}{2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \|\boldsymbol{v}\|^r \psi(\boldsymbol{v} + \hat{\boldsymbol{\theta}}_n) \sup\{\nu_n(\boldsymbol{t}) \colon \|\boldsymbol{t}\| \geqq \delta^* n^{1/2}\}\, d\boldsymbol{v} \\ &\leqq 2^{r-1} n^{\frac{r+1}{2}} \exp - n\,\varepsilon(\delta^*) \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \|\boldsymbol{v}\|^r \psi(\boldsymbol{v})\, d\boldsymbol{v} + \|\hat{\boldsymbol{\theta}}_n\|^r \right] \to 0. \quad \| \end{aligned} \tag{2.35}$$

We complete the proof of the theorem. Again by Taylor's formula,

$$\begin{aligned} \log \nu_n(t) &= \sum_{i=1}^{n} \operatorname{grad}\Phi(z_i, \hat{\boldsymbol{\theta}}_n) \cdot [n^{-1/2}\boldsymbol{t}]' \\ &\quad + \frac{n^{-1}}{2} \sum_{i=1}^{n} \int_0^1 \boldsymbol{t} A(z_i, \hat{\boldsymbol{\theta}}_n + \lambda \boldsymbol{t} n^{-1/2})\, \boldsymbol{t}'\, d\lambda \\ &\leqq \frac{n^{-1}}{2} \sum_{i=1}^{n} \sup\{\boldsymbol{t} A(z_i, \boldsymbol{s})\, \boldsymbol{t}' \colon \|\boldsymbol{s} - \hat{\boldsymbol{\theta}}_n\| \leqq \delta^*\} \\ &\leqq K_1 + \frac{n^{-1}}{2} \sum_{i=1}^{n} \sup\{\boldsymbol{t} A(z_i, \boldsymbol{s})\, \boldsymbol{t}' \colon \|\boldsymbol{s} - \boldsymbol{\theta}_0\| \leqq 2\delta^*\} \end{aligned} \tag{2.36}$$

where K_1 may depend on $z_1, \ldots, z_n, \ldots$. The first inequality follows from (2.17) and the second from (2.16).

Now,

$$\begin{aligned} \sup\{\boldsymbol{t} A(z_1, \boldsymbol{s})\, \boldsymbol{t}' \colon \|\boldsymbol{s} - \boldsymbol{\theta}_0\| \leqq 2\delta^*\} &\leqq \boldsymbol{t} A(z_1, \boldsymbol{\theta}_0)\, \boldsymbol{t}' \\ &+ \|\boldsymbol{t}\|^2 \sup\{\boldsymbol{t}[A(z_1, \boldsymbol{s}) - A(z_1, \theta_0)]\, \boldsymbol{t}' \colon \|\boldsymbol{s} - \boldsymbol{\theta}_0\| \leqq 2\delta^*, \|\boldsymbol{t}\| = 1\}. \end{aligned} \tag{2.37}$$

But,

$$\sup\{\boldsymbol{t}[A(z_1, \boldsymbol{s}) - A(z_1, \boldsymbol{\theta}_0)]\, \boldsymbol{t}' \colon \|\boldsymbol{s} - \boldsymbol{\theta}_0\| < 2\delta^*, \|\boldsymbol{t}\| = 1\} \to 0 \tag{2.38}$$

as $\delta^* \to 0$ by A 2.5.

On the other hand A 2.6 guarantees that for some $\delta^* > 0$ the left hand side of (2.38) which is bounded in absolute value by,

$$\sum_{1 \leqq j, r \leqq k} \sup\left\{\left|\frac{\partial\Phi(\boldsymbol{s}, z_1)}{\partial\theta_j\, \partial_r}\right| \colon \|\boldsymbol{s} - \boldsymbol{\theta}_0\| \leqq \delta^*\right\}$$

has a finite expectation, and hence the expected value of the left hand side of (2.38) tends to 0 also. Then, (2.36), (2.37), the preceding remark and the S.L.L.N. lead us to conclude that for every $\varepsilon > 0$, there exists a $\delta^*(\varepsilon)$ and

$$N^*(\delta^*, z_1, \ldots, z_n, \ldots) < \infty$$

such that,

$$\begin{aligned}\log \nu_n(\boldsymbol{t}) \leqq \boldsymbol{t} A(\boldsymbol{\theta}_0)\boldsymbol{t}' + \varepsilon\|\boldsymbol{t}\|^2 \\ + \|\boldsymbol{t}\|^2 \sup\left\{\boldsymbol{t}\left(n^{-1}\sum_{i=1}^{n} A(z_i, \boldsymbol{\theta}_0) - A(\boldsymbol{\theta}_0)\right)\boldsymbol{t}' : \|\boldsymbol{t}\| = 1\right\}.\end{aligned} \tag{2.39}$$

Yet another application of the S.L.L.N. shows that if I is the identity matrix,

$$\log \nu_n(\boldsymbol{t}) \leqq \boldsymbol{t}(A(\boldsymbol{\theta}_0) - 2\varepsilon I)\boldsymbol{t}' \tag{2.40}$$

for n sufficiently large possibly dependent on the sample sequence and $\boldsymbol{\theta}_0$ but not $\boldsymbol{t}$. By A 2.8 $A(\boldsymbol{\theta}_0) - \varepsilon I$ is negative definite for ε sufficiently small. We conclude that $(1 + \|\boldsymbol{t}\|)^r \nu_n(\boldsymbol{t})$ is bounded by a Lebesgue integrable function, viz. $(1 + \|\boldsymbol{t}\|)^r \exp \boldsymbol{t}(A(\boldsymbol{\theta}_0) - \varepsilon I)\boldsymbol{t}'$ for n sufficiently large, δ^* sufficiently small and all $\boldsymbol{t}$ with $\|\boldsymbol{t} n^{-1/2}\| \leqq \delta^*$. Clearly then, $(1 + \|\boldsymbol{t}\|^r) H(\boldsymbol{t}, z_1, \ldots, z_n)$ is similarly bounded. But $(1 + \|\boldsymbol{t}\|^r) H(t, z_1, \ldots, z_n) \to 0$ by the S.L.L.N. Finally, (2.31) follows by the dominated convergence theorem and theorem 2.2 is proved.

We proceed to the proof of theorem 2.3. The key to the argument is, (compare theorem 10(2) of [12]).

Lemma 2.7. *If* A 2.1 – A 2.9 *hold then*

$$n^{1/2}[\hat{g}_n(z_1, \ldots, z_n) - g(\hat{\boldsymbol{\theta}}_n)] \to 0\,. \tag{2.41}$$

Proof. We begin by proving the weaker,

$$\hat{g}_n - g(\boldsymbol{\theta}_0) \to 0\,. \tag{2.42}$$

We remark that, by (2.8) a)

$$\begin{aligned} Y_n &\leqq \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \tilde{l}(|g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}_0)|)\,\psi(\boldsymbol{\theta} \mid z_1, \ldots, z_n)\, d\boldsymbol{\theta} \\ &= \int_{[\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leqq \delta^*]} \tilde{l}(|g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}_0)|)\,\psi(\boldsymbol{\theta} \mid z_1, \ldots, z_n)\, d\boldsymbol{\theta} \\ &\quad + \int_{[\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \delta^*]} \tilde{l}(|g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}_0)|)\,\psi(\boldsymbol{\theta} \mid z_1, \ldots, z_n)\, d\boldsymbol{\theta}. \end{aligned} \tag{2.43}$$

We can choose $\delta^*(\varepsilon)$ so that by continuity of g and $\tilde{l}$ the first term on the right hand side of (2.43) is $< \varepsilon$. By (2.2) and (2.4) the second term is bounded by

$$\int_{[\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \delta^*]} [K_1 + K_2\|\boldsymbol{\theta}\|^r]\,\psi(\boldsymbol{\theta} \mid z_1, \ldots, z_n)]\, d\boldsymbol{\theta}$$

and therefore tends to 0 by (2.29). Thus

$$Y_n \to 0\,. \tag{2.44}$$

We argue that along every sample sequence satisfying (2.44) and (2.22), (2.42) must hold. For suppose without loss of generality that $\liminf(\hat{g}_n - g(\boldsymbol{\theta}_0)) \geqq \varepsilon > 0$.

Then,

$$\begin{aligned}\liminf_n Y_n &\geqq \liminf_n \int_{[\hat g_n > g(\boldsymbol\theta)]} \tilde l(\hat g_n - g(\boldsymbol\theta))\,\psi(\boldsymbol\theta\,|\,z_1,\ldots,z_n)\,d\boldsymbol\theta \\ &\geqq \tilde l\left(\frac{\varepsilon}{2}\right)\liminf_n \int_{[g(\boldsymbol\theta)<g(\boldsymbol\theta_0)+\varepsilon/2]} \psi\boldsymbol\theta(\,|\,z_1,\ldots,z_n)\,d\boldsymbol\theta = \tilde l\left(\frac{\varepsilon}{2}\right) > 0\,.\end{aligned} \tag{2.45}$$

The last identity is a consequence of the continuity of g and theorem 2.2. We now establish that if (2.22) and (2.34) are satisfied for a sample sequence, (2.41) must hold. For such a sample sequence, by (2.42), (2.2) and (2.4),

$$\begin{aligned}&\int_{[|\hat g_n - g(\boldsymbol\theta_0)|\geqq\delta^*]} \tilde l'(|g(\boldsymbol\theta) - \hat g_n|)\,\psi(\boldsymbol\theta\,|\,z_1,\ldots,z_n)\,d\boldsymbol\theta \\ &\qquad\leqq K \int_{[|g(\boldsymbol\theta)-g(\boldsymbol\theta_0)|\geqq\delta^*]} (1 + \|\boldsymbol\theta\|^r)\,\psi(\boldsymbol\theta\,|\,z_1,\ldots,z_n)\,d\boldsymbol\theta\end{aligned} \tag{2.46}$$

for n sufficiently large. From the continuity of g and (2.34) we see that the right hand side of (2.46) is $0(n^{-\alpha})$ for every $\alpha > 0$. An elementary argument involving (2.2), (2.3), (2.9) and (2.46) leads to,

$$\begin{aligned}&\int_{[g(\boldsymbol\theta)<\hat g_n]} |g(\boldsymbol\theta) - \hat g_n|^s\,\psi(\boldsymbol\theta\,|\,z_1,\ldots,z_n)\,d\boldsymbol\theta \\ &- \int_{[g(\boldsymbol\theta)>\hat g_n]} |g(\boldsymbol\theta) - \hat g_n|^s\,\psi(\boldsymbol\theta\,|\,z_1,\ldots,z_n)\,d\boldsymbol\theta \\ &= 0\{\max(n^{-\alpha}, \textstyle\int |g(\boldsymbol\theta) - \hat g_n|^s\,\psi(\boldsymbol\theta\,|\,z_1,\ldots,z_n)\,d\boldsymbol\theta)\}\end{aligned} \tag{2.47}$$

for every $\alpha > 0$, for the specified sample sequences.

Let us denote by $\mathfrak{L}(h(\boldsymbol\theta)\,|\,z_1,\ldots,z_n)$ the probability law of $h(\boldsymbol\theta)$ if $\boldsymbol\theta$ is distributed according to the density $\psi(\boldsymbol\theta\,|\,z_1,\ldots,z_n)$. Then, lemma 5 of [14] and, a fortiori, theorem 2.2 imply that, for sequences $\{z_i\}$ satisfying (2.22),

$$\mathfrak{L}(n^{1/2}(\boldsymbol\theta - \hat{\boldsymbol\theta}_n)\,|\,z_1,\ldots,z_n) \to G(-A^{-1}(\boldsymbol\theta_0), 0) \tag{2.48}$$

where $G(B, \boldsymbol t)$ is the law of the k variate normal distribution with covariance matrix B and mean $\boldsymbol t$. By a standard argument (for example Cramer [5] p. 366),

$$\mathfrak{L}(n^{1/2}(g(\boldsymbol\theta) - g(\hat{\boldsymbol\theta}_n))\,|\,z_1,\ldots,z_n) \to \mathfrak{N}(0, \sigma^2(g, \boldsymbol\theta_0))\,. \tag{2.49}$$

Suppose without loss of generality that $\lim_n \hat g_n = g(\boldsymbol\theta_0) + c$, $0 < c \leqq \infty$, for a sequence satisfying (2.22), (2.34) and hence (2.47).

Suppose $c < \infty$. We can then conclude, for any $M < \infty$,

$$\begin{aligned}&\int_{[0\leqq n^{1/2}(g(\boldsymbol\theta)-\hat g_n)\leqq M]} n^{s/2}|g(\boldsymbol\theta) - \hat g_n|^s\,\psi(\boldsymbol\theta\,|\,z_1,\ldots,z_n)\,d\boldsymbol\theta \\ &\qquad\to \sigma^{-1}(g, \boldsymbol\theta_0)\int_c^{M+c} \varphi(t\,\sigma^{-1}(g, \boldsymbol\theta_0))\,|t - c|^s\,dt\end{aligned} \tag{2.50}$$

where φ is the standard normal density. But, for any $\varepsilon > 0$, there exists M such that,

$$\int_{[(g(\boldsymbol\theta)-g(\hat{\boldsymbol\theta}_n))\geqq Mn^{-1/2}]} n^{s/2}|g(\boldsymbol\theta) - \hat g_n|^s\,\psi(\boldsymbol\theta\,|\,z_1,\ldots,z_n)\,d\boldsymbol\theta \leqq \varepsilon\,. \tag{2.51}$$

To see this note that by A 2.3 if $\|\boldsymbol\theta - \boldsymbol\theta_0\| < \delta^*$ sufficiently small

$$|g(\boldsymbol\theta) - g(\hat{\boldsymbol\theta}_n)| \leqq K(\delta^*)\,\|\boldsymbol\theta - \hat{\boldsymbol\theta}_n\|\,. \tag{2.52}$$

Therefore,

$$\limsup_n \int_{[(g(\boldsymbol{\theta})-g(\hat{\boldsymbol{\theta}}_n))\geqq Mn^{-1/2}]} n^{s/2}|g(\boldsymbol{\theta})-\hat{g}_n|^s\psi(\boldsymbol{\theta}|z_1,\dots,z_n)\,d\boldsymbol{\theta}$$
$$\leqq \limsup_n \int_{[n^{1/2}\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_n\|\geqq MK^{-1}(\delta^*)]} \|\,n^{1/2}(\boldsymbol{\theta}-\boldsymbol{\theta}_n)\|^s d\boldsymbol{\theta}$$
$$+\int_{[\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n\|\geqq\delta^*]} n^{s/2}|g(\boldsymbol{\theta})-\hat{g}_n|^s\psi(\boldsymbol{\theta}|z_1,\dots,z_n)\,d\boldsymbol{\theta}. \tag{2.53}$$

For M sufficiently large δ^* fixed the first term on the right of (2.53) is $\leqq \varepsilon$ by (2.22) while the second term goes to 0 by (2.2) and (2.34). (2.51) and (2.50) imply by easy further arguments that,

$$\int_{[g(\boldsymbol{\theta})<\hat{g}_n]} n^{s/2}|g(\boldsymbol{\theta})-\hat{g}_n|^s\psi(\boldsymbol{\theta}|z_1,\dots,z_n)\,d\boldsymbol{\theta}$$
$$\to\int_{-\infty}^{c}\sigma^{-1}(g,\boldsymbol{\theta}_0)\,\varphi(t\,\sigma^{-1}(g,\boldsymbol{\theta}_0))\,|t-c|^s\,dt, \tag{2.54}$$

$$\int_{[g(\boldsymbol{\theta})>\hat{g}_n]} n^{s/2}|g(\boldsymbol{\theta})-\hat{g}_n|^s\psi(\boldsymbol{\theta}|z_1,\dots,z_n)\,d\boldsymbol{\theta}$$
$$\to\int_{c}^{\infty}\sigma^{-1}(g,\boldsymbol{\theta}_0)\,\varphi(t\,\sigma^{-1}(g,\boldsymbol{\theta}_0))\,|t-c|^s\,dt, \tag{2.55}$$

and

$$\lim_n n^{s/2}\int|g(\boldsymbol{\theta})-\hat{g}_n|^s\psi(\boldsymbol{\theta}|z_1,\dots,z_n)\,d\boldsymbol{\theta}>0. \tag{2.56}$$

But

$$\int_{-\infty}^{c}|t-c|^s\varphi(at)\,dt \neq \int_{c}^{\infty}|t-c|^s\varphi(at)\,dt \tag{2.57}$$

unless $c=0$.

Thus, (2.47) is contradicted by (2.54)—(2.57) unless $c=0$.

The case $c=\pm\infty$ follows by similar arguments. ||

We have proved part a) of theorem 2.3 in the lemma. The lemma implies that $n^{1/2}(\hat{g}_n-g(\boldsymbol{\theta}_0))$ has the same asymptotic behavior as $n^{1/2}(g(\hat{\boldsymbol{\theta}}_n)-g(\boldsymbol{\theta}_0))$. But,

$$\mathfrak{L}_{\boldsymbol{\theta}_0}(n^{1/2}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0))\to G(-A^{-1}(\boldsymbol{\theta}_0),0) \tag{2.58}$$

by standard arguments yielding asymptotic normality of maximum likelihood estimates c.f. ([15] pp. 500—506), and part b) of the theorem again follows by the classical Taylor expansion Slutsky theorem argument of (say) [5] p. 366. Finally, we complete the proof of theorem 2.4.

$$n^\beta Y_n = n^\beta\int_{[n^{1/2}(g(\boldsymbol{\theta})-\hat{g}_n)\leqq M]}\tilde{l}(|g(\boldsymbol{\theta})-\hat{g}_n|)\,\psi(\boldsymbol{\theta}|z_1,\dots,z_n)\,d\boldsymbol{\theta}$$
$$+\,n^\beta\int_{[n^{1/2}(g(\boldsymbol{\theta})-\hat{g}_n)>M]}\tilde{l}(|g(\boldsymbol{\theta})-\hat{g}_n|)\,\psi(\boldsymbol{\theta}|z_1,\dots,z_n)\,d\boldsymbol{\theta}. \tag{2.59}$$

The second term may be shown to be $\leqq\varepsilon$ for M sufficiently large and all n by an argument similar to that leading to (2.51) since by (2.3) there exists δ^* such that

$$\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n\|\leqq\delta^*\Rightarrow|g(\boldsymbol{\theta})-g(\hat{\boldsymbol{\theta}}_n)|<\delta^{**}$$
$$\Rightarrow\tilde{l}(|g(\boldsymbol{\theta})-g(\hat{\boldsymbol{\theta}}_n)|)\leqq K|g(\boldsymbol{\theta})-g(\hat{\boldsymbol{\theta}}_n)|^{s+1}.$$

Again applying (2.3), (2.49) and (2.59) in arguments similar to those used to establish (2.54) and (2.55) we see that,

$$n^\beta Y_n \sim \gamma \int_n^{(s+1)/2} |g(\boldsymbol{\theta}) - g(\hat{\boldsymbol{\theta}}_n)|^{s+1} \psi(\boldsymbol{\theta} | z_1, \dots, z_n)\, d\boldsymbol{\theta} \tag{2.60}$$

$$\sim \sigma^{-1}(g, \boldsymbol{\theta}_0) \gamma \int_{-\infty}^{\infty} |t|^{2\beta} \varphi(\sigma^{-1}(g, \boldsymbol{\theta}_0) t)\, dt.$$

Theorem 2.4 is proved.

Proof of Theorem 2.5. We prove the theorem for the special case $k = 1$, $g(\theta) = \theta$, leaving the rather tedious details of the general case to the reader. Now,

$$n^\beta Y_n - V(\boldsymbol{\theta}) = \gamma(s+1)^{-1} \int_{-\infty}^{\infty} |t|^{s+1} \{\psi^g(t | z_1, \dots, z_n) \tag{2.61}$$

$$- \sigma^{-1}(g, \theta) \varphi(t[\sigma(g, \theta)]^{-1})\}\, dt$$

where

$$\psi^g(t | z_1, \dots, z_n) = \left[\prod_{i=1}^n f(z_i, t n^{-1/2} + \hat{g}_n)\right] \psi(t n^{-1/2} + g_n) \tag{2.62}$$

$$\cdot \left[\int_{-\infty}^{\infty} \prod_{i=1}^n f(z_i, s n^{-1/2} + \hat{g}_n) \psi(s n^{-1/2} + \hat{g}_n)\, ds\right]^{-1}.$$

Define,

$$\gamma_n^g(t) = \sum_{i=1}^n [\Phi(z_i, t n^{-1/2} + \hat{g}_n) - \Phi(z_i, \hat{\theta}_n)] \tag{2.63}$$

where $\hat{\theta}_n = \hat{\boldsymbol{\theta}}_n$ for the case $k = 1$.

Consider

$$T_n^{(1)} = \int_{-\infty}^{\infty} |t|^{s+1} \left\{\exp \gamma_n^g(t) - \exp \frac{-1}{2} t^2 \sigma^{-2}(g, \theta)\right\} dt \tag{2.64}$$

and

$$T_n^{(2)} = \int_{-\infty}^{\infty} \left\{\exp \gamma_n^g(t) - \exp \frac{-1}{2} t^2 \sigma^{-2}(g, \theta)\right\} dt. \tag{2.65}$$

Note that $\sigma^{-2}(g, \theta)$ corresponds to $-A(\boldsymbol{\theta})$.

We remark if $u_n \to u_0$, $v_n \to v_0$

$$\begin{aligned} u_n v_n^{-1} &= u_0 v_0^{-1} + v_n^{-1} \{(u_n - u_0) + u_0 v_0^{-1} (v_0 - v_n)\} \\ &= u_0 v_0^{-1} + v_0^{-1}\{(u_n - u_0) + u_0 v_0^{-1}(v_0 - v_n)\} \\ &\quad + 0((v_n - v_0) v_0^{-1}) [(u_n - u_0) + (v_n - v_0)]\,. \end{aligned} \tag{2.66}$$

As a consequence we see that the theorem follows if we show

$$T_n^{(1)} = \left[\frac{1}{2} \int_{-\infty}^{\infty} |t|^{s+3} \exp \frac{-t^2}{2} \sigma^{-2}(g, \theta)\, dt\right]$$

$$\left\{n^{-1} \sum_{i=1}^n [A(z_i, \theta) - E_\theta(A(z_i, \theta)]\right. \tag{2.67}$$

$$\left. + E_\theta\left(\frac{\partial A(z_1, \theta)}{\partial \theta}\right) E_\theta^{-1}(A(z_1, \theta)) n^{-1} \sum_{i=1}^n \frac{\partial \Phi(z_i, \theta)}{\partial \theta}\right\} + 0(n^{-1/2}),$$

$$(2.68)\quad T_n^{(2)} = \left[\frac{1}{2}\int_{-\infty}^{\infty} t^2 \exp \frac{-t^2}{2}\sigma^{-2}(g,\theta)\right]\Big\} n^{-1}\sum_{i=1}^{n}[A(z_i,\theta) - E_\theta(A(z_1,\theta)]$$
$$+ E_\theta\left(\frac{\partial A(z_i,\theta)}{\partial\theta}\right) E_\theta^{-1}[(A(z_1,\theta))\, n^{-1}\sum_{i=1}^{n}\frac{\partial\Phi(z_1,\theta)}{\partial\theta}\Big\} + 0(n^{-1/2}).$$

By arguments similar to those given in lemma 2.5 we see that,

$$(2.69)\qquad T_n^{(1)} = \int_{|t|\leq s n^{1/2}} |t|^{s+1}\{\exp\gamma_n^g(t) - \exp - \tfrac{1}{2} t^2\sigma^{-2}(g,\theta)\}\,dt + 0(n^{-\alpha})$$

a.s. P_θ for all $\alpha > 0$ all $\delta > 0$ sufficiently small and that a similar statement holds for $T_n^{(2)}$.

Now

$$(2.70)\quad \exp\gamma_n^g(t) = \exp - \frac{t^2}{2}\sigma^{-2}(g,\theta) + [\exp\lambda_n(t)]\,(\gamma_n^g(t) + \frac{t^2}{2}\sigma^{-2}(g,\theta))$$

where $\lambda_n(t)$ lies between $\gamma_n^g(t)$ and $-t^2/2\,\sigma^{-2}(g,\theta)$. But,

$$(2.71)\qquad \gamma_n^g(t) = \frac{t^2}{2}\sum_{i=1}^{n}\int_0^1 \lambda A(z_i, \lambda(t n^{-1/2} + (\hat g_n - \hat\theta_n)) + \hat\theta_n)\,d\lambda.$$

Thus,

$$(2.72)\qquad \begin{aligned}\gamma_n^g(t) + \frac{t^2}{2}\sigma^{-2}(g,\theta) &= \frac{t^2}{2} n^{-1}\sum_{i=1}^{n}[A(z_i,\theta) + \sigma^{-2}(g,\theta)]\\ &+ t^2 n^{-1}\sum_{i=1}^{n}\int_0^1 \lambda[A(z_i, \lambda(t n^{-1/2} + (\hat g_n - \hat\theta_n)) + \hat\theta_n) - A(z_i,\theta)]\,d\lambda.\end{aligned}$$

Expanding $A(z,\theta)$ we get by A 2.6′

$$(2.73)\qquad \begin{aligned}&A(z_i, \lambda(t n^{-1/2} + (\hat g_n - \hat\theta_n) + \hat\theta_n) - A(z_i,\theta))\\ &= \left[\frac{\partial}{\partial\theta}A(z_i, \gamma_n(\lambda)(t n^{-1/2} + (\hat g - \hat\theta_n)) + \hat\theta_n)\right]\{t n^{-1/2} + (\hat g_n - \hat\theta_n)\}\lambda\\ &+ [A(z_i,\hat\theta_n) - A(z_i,\theta)]\end{aligned}$$

where $\gamma_n(\lambda)$ lies between 0 and λ. Using A 2.6′ and A 2.7′ and the dominated convergence theorem as in the proof of theorem 2.2 it is easy to see that,

$$(2.74)\qquad \begin{aligned}&n^{1/2}\int_{|t|\leq\delta n^{1/2}} |t|^{s+1}(t n^{-1/2} + (\hat g_n - \hat\theta_n))\,\lambda n^{-1} t^2\\ &\cdot\sum_{i=1}^{n}\int_0^1 \lambda^2\frac{\partial A}{\partial\theta}[z_i, \gamma_n(\lambda)(t n^{-1/2} + (\hat g_n - \hat\theta_n)) + \hat\theta_n]\,d\lambda\exp\lambda_n(t)\,dt\\ &- \frac{1}{3}\int_{-\infty}^{\infty}|t|^{s+3}[t + n^{1/2}(\hat g_n - \hat\theta_n)]\,E_\theta\left(\frac{\partial A(z_1,\theta)}{\partial\theta}\right)\\ &\cdot\left[\exp - \frac{t^2}{2}\sigma^{-2}(g,\theta)\right]dt \to 0.\end{aligned}$$

But the last term on the right hand side of (2.74) tends to 0 by lemma 2.6.

In conclusion,

$$(2.75)\quad \begin{aligned} n^{-1}\sum_{i=1}^{n}[A(z_i,\hat\theta_n)-A(z_i,\theta)] &= \left\{n^{-1}\sum_{i=1}^{n}\frac{\partial A(z_i,\theta)}{\partial\theta}\right\}(\hat\theta_n-\theta) \\ &+\frac{n^{-1}}{2}\sum_{1}^{n}\int_0^1\frac{\partial^2 A(z_i,\theta+\lambda(\hat\theta_n-\theta))}{\partial\theta^2}d\lambda(\hat\theta_n-\theta)^2 \\ &= n^{-1}\sum_{i=1}^{n}\frac{\partial A(z_1,\theta)}{\partial\theta}(\hat\theta_n-\theta)+0(\hat\theta_n-\theta)^2 \end{aligned}$$

by A 2.7′.

But

$$(2.76)\quad (\hat\theta_n-\theta)=\left\{n^{-1}\sum_{i=1}^{n}\frac{\partial\Phi(z_i,\theta)}{\partial\theta}\right\}\left\{n^{-1}\sum_{i=1}^{n}A(z_i,\hat\theta_n)\right\}^{-1}+0(\hat\theta_n-\theta)^2$$

by the likelihood equation and A 2.7′.

Using the law of the iterated logarithm and 2.7 b) we get,

$$(2.77)\quad (\hat\theta_n-\theta)^2=0(n^{-1}[\log\log n]),$$

$$(2.78)\quad n^{-1}\sum_{i=1}^{n}\frac{\partial A(z_1,\theta)}{\partial\theta}=E_\theta\left(\frac{\partial A(z_i,\theta)}{\partial\theta}\right)+0\left[\frac{\log\log n}{n}\right]^{1/2}.$$

Upon using (2.75) we also have,

$$(2.79)\quad n^{-1}\sum_{i=1}^{n}A(z_i,\hat\theta_n)=-\sigma^{-2}(g,\theta)+0\left[\frac{\log\log n}{n}\right]^{1/2}.$$

Finally we obtain

$$(2.80)\quad \begin{aligned} &n^{-1}\sum_{i=1}^{n}[A(z_i,\hat\theta_n)-A(z_i,\theta)] \\ &\quad=-E_\theta\left(\frac{\partial A(z_i,\theta)}{\partial\theta}\right)\sigma^2(g,\theta)\,n^{-1}\sum_{i=1}^{n}\frac{\partial\Phi(z_i,\theta)}{\partial\theta}+0(n^{-1/2}) \end{aligned}$$

since by the law of the interated logarithm

$$n^{-1}\sum_{i=1}^{n}\frac{\partial\Phi(z_i,\theta)}{\partial\theta}=0\left[\frac{\log\log n}{n}\right]^{1/2}.$$

Then (2.67) follows from (2.80) and (2.72)—(2.73). A similar argument yields (2.68) and the theorem follows. ||

3. Testing

Our assumptions throughout this section are much less restrictive than those of section 2. They may be weakened even further and put essentially in the form given in [2]. We feel the present level of generality is adequate for most purposes. Admittedly, difficulties of the same nature as those arising in the application of the Wald [16] conditions to the normal model do occur but they may be dealt with by devices such as those of [10].

As before we suppose the z_i have a density $f(z, \boldsymbol{\theta})$ with respect to μ. We continue to assume that Θ is an open subset of R^k endowed with its usual topology.

However, we drop the requirement that our prior measure Ψ has a density with respect to Lebesgue measure, but ask merely that Ψ assign positive mass to every nonempty open subset of Θ.

Our decision space D now contains two members d_0 and d_1. We suppose that for every $\boldsymbol{\theta}$ at least one of $l(\boldsymbol{\theta}, d_0)$, $l(\boldsymbol{\theta}, d_1)$ equals 0, and both loss functions are measurable in $\boldsymbol{\theta}$ and nonnegative.

Let,

$$l(\boldsymbol{\theta}) = \max[l(\boldsymbol{\theta}, d_1), l(\boldsymbol{\theta}, d_0)]. \tag{3.1}$$

We suppose that,

$$\int_{\Theta} l(\boldsymbol{\theta}) \Psi(d\boldsymbol{\theta}) < \infty \tag{3.2}$$

and for convenience in notation define a measure Ψ^* on $\mathfrak{B}$ by

$$\Psi^*(B) = \int_B l(\boldsymbol{\theta}) \Psi(d\boldsymbol{\theta}). \tag{3.3}$$

Our hypothesis is given by the set,

$$H = \{\boldsymbol{\theta} : l(\boldsymbol{\theta}, d_1) = l(\boldsymbol{\theta})\}. \tag{3.4}$$

Thus H includes the indifference region if any exists.

A Bayes procedure is given by,

$$\delta(z_1, \ldots, z_n) = d_0 \tag{3.5}$$

if

$$\int_H \prod_{i=1}^n f(z_i, \boldsymbol{\theta}) \Psi^*(d\boldsymbol{\theta}) \geqq \int_{H'} \prod_{i=1}^n f(z_i, \boldsymbol{\theta}) \Psi^*(d\boldsymbol{\theta}), \quad \delta(z_1, \ldots, z_n) = d_1$$

otherwise. H' denotes the complement of H as usual. The Bayes posterior risk is,

$$\begin{aligned} Y_n = \min \Bigg\{ & \int_H \prod_{i=1}^n f(z_i, \boldsymbol{\theta}) \Psi^*(d\theta), \\ & \cdot \int_{H'} \prod_{i=1}^n f(z_i, \boldsymbol{\theta}) \Psi^*) (d\boldsymbol{\theta}) \Bigg\} \left[\int_{\Theta} \prod_{i=1}^n f(z_i, \boldsymbol{\theta}) \Psi(d\boldsymbol{\theta}) \right]^{-1}. \end{aligned} \tag{3.6}$$

The main assumptions of this section are the following,

A 3.1. $\Psi(U) > 0$ for all open U, and

$$0 < \Psi^*(H) < \Psi^*(\Theta). \tag{3.7}$$

A 3.2. $\Phi(z_1, \boldsymbol{\theta})$ is separable in the sense of Doob when considered as a process in $\boldsymbol{\theta}$.

A 3.3. Define,

$$h(\boldsymbol{\theta}_0, \delta, \boldsymbol{s}) = E_{\boldsymbol{\theta}_0}[\sup\{|\Phi(\boldsymbol{z}_1, s) - \Phi(\boldsymbol{z}_1, \boldsymbol{t})| : \|\boldsymbol{s} - \boldsymbol{t}\| \leqq \delta\}]. \tag{3.8}$$

We require that for δ sufficiently small h is finite and,

$$\lim_{\delta \to 0} h(\boldsymbol{\theta}_0, \delta, \boldsymbol{s}) = 0. \tag{3.9}$$

A 3.4. For some $d(\boldsymbol{\theta}_0) < \infty$,

$$E_{\boldsymbol{\theta}_0}[\sup\{[\Phi(z_1, \boldsymbol{\theta}) - \Phi(z_1, \boldsymbol{\theta}_0)] : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geqq d(\boldsymbol{\theta}_0)\}] \leqq B(\boldsymbol{\theta}_0) \tag{3.10}$$

where $B(\boldsymbol{\theta}_0)$ is defined below. We can now state a better version of theorem 4.2 of [3].

Theorem 3.1. *If assumptions* A 3.1—A 3.4 *hold*

$$n^{-1}\log Y_n \to B(\boldsymbol{\theta}_0), \tag{3.11}$$

where

$$B(\boldsymbol{\theta}_0) = \min\{(\Psi^*)\operatorname{ess\,sup}\{J(\boldsymbol{\theta},\boldsymbol{\theta}_0):\boldsymbol{\theta}\in H\}, (\Psi^*)\operatorname{ess\,sup}\{J(\boldsymbol{\theta},\boldsymbol{\theta}_0):\boldsymbol{\theta}\in H'\}\}. \tag{3.12}$$

$$J(\boldsymbol{\theta},\boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}_0}[\Phi(z_1,\boldsymbol{\theta}) - \Phi(z_1,\boldsymbol{\theta}_0)] \tag{3.13}$$

and $(\Psi^*)\operatorname{ess\,sup}$ *denotes the essential supremum of the quantity within brackets with respect to* Ψ^* *measure.* (This definition corrects an error made in the definition of $B(\boldsymbol{\theta}_0)$ in [3].)

The proof we shall give is essentially a "two sided" version of the proof of lemma 3 of section 4 in [2]. Our original proof of theorem 4.2 of [3] did not generalize readily when assumption (4) and (5) of [3] were weakened to A 3.2 and A 3.3. We are indebted to R. R. BAHADUR for pointing out that A 3.3 sufficed for our result.

Proof of theorem 3.1. We define,

$$J_n(\boldsymbol{\theta},\boldsymbol{\theta}_0) = n^{-1}\sum_{i=1}^{n}[\Phi(z_i,\boldsymbol{\theta}) - \Phi(z_i,\boldsymbol{\theta}_0)]. \tag{3.14}$$

Then,

$$n^{-1}\log Y_n = \min\left\{n^{-1}\log\int_H \exp n J_n(\boldsymbol{\theta},\boldsymbol{\theta}_0)\Psi^*(d\boldsymbol{\theta}),\ n^{-1}\log\int_{H'}\exp n J_n(\boldsymbol{\theta},\boldsymbol{\theta}_0)\Psi^*(d\boldsymbol{\theta})\right\}. \tag{3.15}$$

We begin by noting that,

$$n^{-1}\log\int_\Theta \exp n J_n(\boldsymbol{\theta},\boldsymbol{\theta}_0)\Psi(d\boldsymbol{\theta}) \to 0 \tag{3.16}$$

a.s. $P_{\boldsymbol{\theta}_0}$.

To prove (3.16) it clearly suffices to show that,

$$n^{-1}\log\int_{\bar S}\exp n J_n(\boldsymbol{\theta},\boldsymbol{\theta}_0)\Psi(d\boldsymbol{\theta}) \to 0 \tag{3.17}$$

and

$$\limsup_n n^{-1}\log\int_{\bar S'}\exp n J_N(\boldsymbol{\theta},\boldsymbol{\theta}_0)\Psi(d\boldsymbol{\theta}) \leqq 0 \tag{3.18}$$

a.s. $P_{\boldsymbol{\theta}_0}$, where $\bar S$ is the set $\{\boldsymbol{\theta}: \|\theta - \theta_0\| \leqq d(\boldsymbol{\theta})\}$ and $'$ denotes complementation. Without loss of generality we may suppose $\bar S$ is disjoint from the boundary of Θ and hence is compact.

Now,

$$\begin{aligned} & n^{-1}\log\int_{\bar S'}\exp n J_n(\boldsymbol{\theta},\boldsymbol{\theta}_0)\Psi(d\boldsymbol{\theta}) \\ & \leqq n^{-1}\sum_{i=1}^{n}\sup\{[\Phi(z_i,\boldsymbol{\theta}) - \Phi(z_i,\boldsymbol{\theta}_0)]: \|\theta-\theta_0\| \geqq d(\boldsymbol{\theta}_0)\} + n^{-1}\log\Psi(\bar S'). \end{aligned} \tag{3.19}$$

Since $\log\Psi(\bar S') \leqq 0$ by the strong law of large numbers (S.L.L.N.), (3.19) and A 3.4 the left hand side of (3.18) is $\leqq B(\boldsymbol{\theta}_0)$ which is, of course, $\leqq 0$.

Now,

$$(\Psi)\operatorname{ess\,sup} J(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \sup_{\theta} J(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = 0 \tag{3.20}$$

by A 3.1 and A 3.3, since A 3.3 clearly implies that J is continuous in $\boldsymbol{\theta}$.

In view of lemma 4.2 of [3], (3.17) will follow if we can show,

$$\sup\{|Q_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|: \boldsymbol{\theta} \in \bar{S}\} \to 0 \tag{3.21}$$

a.s. $P_{\boldsymbol{\theta}_0}$ where,

$$Q_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = J_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - J(\boldsymbol{\theta}, \boldsymbol{\theta}_0). \tag{3.22}$$

By A 3.3 given $\varepsilon > 0$, there exist $\delta(\varepsilon, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ such that,

$$h(\boldsymbol{\theta}_0, \delta, \boldsymbol{\theta}) < \varepsilon. \tag{3.23}$$

Cover $\bar{S}$ by putting about each $\boldsymbol{\theta}$ a sphere of radius $\delta(\varepsilon, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$. Extract a finite subcovering centered at (say) $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r$

$$\begin{aligned}
&\sup\{|Q_n(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - Q_n(\boldsymbol{\theta}_j, \boldsymbol{\theta}_0)|: \|\boldsymbol{\theta}_j - \boldsymbol{\theta}\| \leqq \delta(\varepsilon, \boldsymbol{\theta}_j, \boldsymbol{\theta}_0)\} \\
&\quad \leqq n^{-1} \sum_{i=1}^{n} \sup\{|\Phi(z_i, \boldsymbol{\theta}) - \Phi(z_i, \boldsymbol{\theta}_j)|: \|\boldsymbol{\theta} - \boldsymbol{\theta}_j\| \leqq \delta(\varepsilon, \boldsymbol{\theta}_j, \boldsymbol{\theta})\} \\
&\quad + \sup\{|J(\boldsymbol{\theta}_j, \boldsymbol{\theta}_0) - J(\boldsymbol{\theta} - \boldsymbol{\theta}_0)|: \|\boldsymbol{\theta} - \boldsymbol{\theta}_j\| \leqq \delta(\varepsilon, \boldsymbol{\theta}_j, \boldsymbol{\theta})\}.
\end{aligned} \tag{3.24}$$

We conclude by A 3.3, (3.24) and the S.L.L.N. that,

$$\limsup_n \sup\{|Q_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - Q_n(\boldsymbol{\theta}_j, \boldsymbol{\theta}_0)|: \|\boldsymbol{\theta}_j - \boldsymbol{\theta}\| \leqq \delta(\varepsilon, \boldsymbol{\theta}_j, \boldsymbol{\theta}_0)\} \leqq 2\varepsilon \tag{3.25}$$

a.s. $P_{\boldsymbol{\theta}_0}$.

But, then,

$$\begin{aligned}
&\limsup_n \sup\{|Q_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|: \boldsymbol{\theta} \in \bar{S}\} \leqq \limsup_n \max_{1 \leqq j \leqq r} |Q_n(\boldsymbol{\theta}_j, \boldsymbol{\theta}_0)| \\
&\qquad + \limsup_n \max_{1 \leqq j \leqq r} \sup\{|Q_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - Q_n(\boldsymbol{\theta}_j, \boldsymbol{\theta}_0)|: \|\boldsymbol{\theta}_j - \boldsymbol{\theta}\| \\
&\qquad \leqq \delta(\varepsilon, \boldsymbol{\theta}_j, \boldsymbol{\theta}_0)\} \leqq 2\varepsilon
\end{aligned} \tag{3.26}$$

by (3.25) and the S.L.L.N. applied to Q_n. (3.26) implies (3.21) and (3.17) and (3.16) follow.

To complete the proof of the theorem we need only imitate the proof of (3.20) in showing that,

$$\mathrm{n}^{-1} \log \int_H \exp n J_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \Psi^*(d\boldsymbol{\theta}) \to (\Psi^*) \operatorname{ess\,sup}\{J(\boldsymbol{\theta}, \boldsymbol{\theta}_0): \boldsymbol{\theta} \in H\} \tag{3.27}$$

and similarly for H'.

The only modification that need be made is replacing $\bar{S}$ by $S^*(\varepsilon)$ where $S^*(\varepsilon)$ is a compact so large that,

$$[S^*(\varepsilon)]' \subset \{\boldsymbol{\theta}: \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geqq d(\boldsymbol{\theta}_0)\} \tag{3.28}$$

and

$$(\Psi^*) \operatorname{ess\,sup}\{J(\boldsymbol{\theta}, \boldsymbol{\theta}_0): \boldsymbol{\theta} \in H\} \leqq (\Psi^*) \operatorname{ess\,sup}\{J(\boldsymbol{\theta}, \boldsymbol{\theta}_0): \boldsymbol{\theta} \in S^*(\varepsilon) \cap H\} + \varepsilon \tag{3.29}$$

and then letting $\varepsilon \to 0$.

We remark that the structure conditions of theorems 4.1 and 4.2 of [3] easily imply A 3.1—A 3.4. Conditions (2), (6) and the second part of (3) of theorem 4.2 of [3] are irrelevant to the satisfaction of A 3.1—A 3.4 but are merely designed to ensure that $-\infty < B(\theta_0) < 0$.

As can be seen from theorem 4.1 of [3] the requirement that Θ be an open set is largely irrelevant and is made to avoid further cluttering up our assumptions. A more elegant but in our opinion less immediately useful formulation of the conditions required for the validity of the theorem may be made in terms of BAHADUR's device of a suitable compactification of Θ. Requirements of this type for related problems of asymptotic theory are given in [1] and [2].

In the course of proving consistency of Bayes procedures under very weak assumptions, L. SCHWARTZ [20] showed that Y_n tends to 0 exponentially. However, her conditions are too weak to yield the existence and identity of the limit of $Y_n^{1/n}$.

4. Concluding Remarks and Generalizations

An easy and interesting extension of the theory of sections 2 and 3 may be made to the case of "improper" priors (see e.g. JEFFREYS [18]).

Suppose that we drop the requirement that $\int_\Theta \psi(\boldsymbol{\theta})\,d\boldsymbol{\theta} < \infty$ and substitute instead.

A′ 2.2. There exists $N(z_1, \ldots, z_n, \ldots)$ with $P_{\boldsymbol{\theta}}[N < \infty] = 1$, such that, $\psi(\boldsymbol{\theta})\prod_{i=1}^{n} f(z_i, \boldsymbol{\theta})$ is a bounded continuous function of $\boldsymbol{\theta}$ and,

$$\int_\Theta \psi(\boldsymbol{\theta})\prod_{i=1}^{n} f(z_i, \boldsymbol{\theta})\,d\boldsymbol{\theta} < \infty \quad \text{for} \quad n \geqq N. \tag{4.1}$$

To match A 2.5 we also require,

A′ 2.5. There exists $N(z_1, \ldots, z_n, \ldots, \ldots)$ such that,

$$\int_\Theta \|\boldsymbol{\theta}\|^r \psi(\boldsymbol{\theta})\prod_{i=1}^{n} f(z_i, \boldsymbol{\theta})\,d\boldsymbol{\theta} \quad \text{for} \quad n \geqq N. \tag{4.2}$$

Then, for $n \geqq N$, we can define the probability density $\psi(\boldsymbol{\theta}\,|\,z_1, \ldots, z_n)$ as in (2.8).

It then follows that if we replace A 2.2 and A 2.5 in the assumptions of theorems 2.2—2.4, by A′ 2.2 and A′ 2.5 these theorems continue to hold. This is an easy consequence of the identity,

$$\begin{aligned}(4.3)\quad & \psi(\boldsymbol{\theta}\,|\,z_1, \ldots, z_{n+N}) \\ &= \prod_{j=1}^{n} f(z_{N+j}, \boldsymbol{\theta})\,\psi(\boldsymbol{\theta}\,|\,z_1, \ldots, z_N) \int_\Theta \left[\prod_{j=1}^{n} f(z_{N+j}, \boldsymbol{\theta})\right] \psi(\boldsymbol{\theta}\,|\,z_1, \ldots, z_N)\,d\boldsymbol{\theta}^{-1}.\end{aligned}$$

In other words we may go through the proofs of our theorems verbatim considering experimentation as starting after time N with prior $\psi(\boldsymbol{\theta}\,|\,z_1, \ldots, z_N)$ for a given sample sequence. A similar generalization holds for testing if we modify A 3.1 suitably.

A generalization of theorem 2.2 which has found some application in situations where one considers sequences of loss functions (see [19]) is the following.

Let h be a continuous function from R^k to R such that,

$$\limsup_{\|\boldsymbol{t}\| \to \infty} |h(\boldsymbol{t})| \, \|\boldsymbol{t}\|^{-r} < \infty$$

for some $r < \infty$.

Then, we have,

Theorem 4.1. *Under assumptions* A 2.1—A 2.2, A 2.5—A 2.9 *and* A 4.1,

$$\int_{\Theta} h(n^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_n)) \, \psi(\hat{\boldsymbol{\theta}} \mid z_1, \ldots, z_n) \, d\boldsymbol{\theta} \to \int_{\Theta} h(\boldsymbol{t}) \, \varphi(-A^{-1}(\boldsymbol{\theta}_0), \boldsymbol{t}) \, d\boldsymbol{t}. \tag{4.4}$$

The proof is essentially the same as that of the seemingly less general theorem 2.2 and we do not give it. It is of interest to note that the proof we give will not be satisfied by anything weaker than A 4.1. A thorough examination will show that what seems to be needed is,

A 4.1′

$$M(\alpha) = \sup\{|h(\alpha \boldsymbol{t}) [h(\boldsymbol{t})]^{-1}| : \boldsymbol{t} \in R^k\} < \infty$$

for every $\alpha > 0$.

But it is easy to see that then,

$$M(\alpha\beta) \leqq M(\alpha) M(\beta) \tag{4.5}$$

and if we define,

$$q(x) = \log M(e^x), \tag{4.6}$$

$$q(x+y) \leqq q(x) + q(y). \tag{4.7}$$

By a classical lemma [8] p. 616, there exists $r < \infty$ such that,

$$\lim_{x \to \infty} \frac{q(x)}{x} = r. \tag{4.8}$$

This is equivalent to A 4.1.

Finally, we note that the results of section 3 may be generalized to finite multiple decision procedures. The case still essentially left open is the behavior of Y_n when $\boldsymbol{\theta}_0$ which is a boundary point of hypothesis and alternative holds. This situation can, we believe, be dealt with by the methods of Sethuraman and Rubin [15].

References

1. Bahadur, R. R.: An optimal property of the likelihood ratio statistic. Proc. fifth Berkeley Sympos. math. Statist. Probability (1965).
2. —, and P. J. Bickel: An optimal property of Bayes' test statistics. To be submitted to Sankhyā (1966).
3. Bickel, P. J., and J. A. Yahav: Asymptotically pointwise optimal procedures in sequential analysis. Proc. fifth Berkeley Sympos. math. Statist. Probability (1965).
4. — — Asymptotically optimal procedures in Bayes and minimax sequential estimation. Ann. math. Statistics **39**, 442—456 (1968).
5. Cramèr, H.: Methods of Mathematical Statics. Princeton Univ. Press 1946.
6. Davis, R. C.: Asymptotic properties of Bayes estimate. Ann. math. Statistics **22**, 8,484 (Abstract) (1951).
7. Dieudonné, J.: Foundations of Modern Analysis. New York: Interscience 1961.
8. Dunford, N., and J. Schwartz: Linear Operators Part I. New York: Interscience 1957.
9. Farrell, R.: Weak limits of sequences of Bayes procedures in estimation theory. Proc. fifth Berkeley Sympos. math. Statist. Probability (1965).

10. KIEFER, J., and J. WOLFOWITZ: Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann. math. Statistics **27**, 887–906 (1956).
11. —, and J. SACKS: Asymptotically optimum sequential inference and design. Ann. math. Statistics **34**, 705–750 (1963).
12. LE CAM, L.: On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. Univ. California Publ. Statist. **1**, 277–330 (1953).
13. — On the asymptotic theory of estimation and testing hypotheses. Proc. third Berkeley Symp. math. Statist. Probability, 129–157 (1955).
14. — Les propriétés asymptotiques des solutions de Bayes. Publ. Inst. Statist. Univ. Paris **7**, 18–35 (1958).
15. RUBIN, H., and J. SETHURAMAN: Probabilities of moderate deviations. Sankhyā **27**, 325–346 (1965).
16. WALD, A.: A note on the consistency of the maximum likelihood estimate. Ann. math. Statistics **20**, 595–601 (1949).
17. WOLFOWITZ, J.: Method of maximum likelihood and the Wald theory of decision functions. Indagationes math. **15**, 114–119 (1953).
18. JEFFREYS, H.: Theory of probability, 3rd Ed. Oxford Univ. Press 1961.
19. GOMBERG, D.: Doctoral dissertation. Univ. of California, Berkeley 1967.
20. SCHWARTZ, L.: On Bayes procedures. Z. Wahrscheinlichkeitstheorie verw. Geb. **4**, 10–26 (1965).
21. ELFVING, G.: Robustness of Bayes decisions against choice of prior. Tech. Report 122, Stanford University 1966.
22. BICKEL, P. J., and J. A. YAHAV: On an asymptotically optimal rule in sequential estimation with quadratic loss. To appear in Ann. Math. Statistics, April 1969.

Professor Dr. P. J. BICKEL
University of California
Dept. of Statistics
Berkeley, Calif. 94720, USA

Professor Dr. J. A. YAHAV
University of Tel Aviv
Dept. of Mathem. Statistics
Tel Aviv, Israel

The Annals of Statistics
1990, Vol. 18, No. 3, 1070–1090

A DECOMPOSITION FOR THE LIKELIHOOD RATIO STATISTIC AND THE BARTLETT CORRECTION—A BAYESIAN ARGUMENT

BY PETER J. BICKEL[1,2] AND J. K. GHOSH[2]

University of California, Berkeley and Indian Statistical Institute

Let $l(\theta) = n^{-1}\log p(x,\theta)$ be the log likelihood of an n-dimensional X under a p-dimensional θ. Let $\hat{\theta}_j$ be the mle under H_j: $\theta^1 = \theta_0^1, \dots, \theta^j = \theta_0^j$ and $\hat{\theta}_0$ be the unrestricted mle. Define T_j as

$$\left[2n\left\{l(\hat{\theta}_{j-1}) - l(\hat{\theta}_j)\right\}\right]^{1/2}\operatorname{sgn}\left(\hat{\theta}^j_{j-1} - \theta^j_0\right).$$

Let $T = (T_1, \dots, T_p)$. Then under regularity conditions, the following theorem is proved: Under $\theta = \theta_0$, T is asymptotically $N(n^{-1/2}a_0 + n^{-1}a, J + n^{-1}\Sigma) + O(n^{-3/2})$ where J is the identity matrix. The result is proved by first establishing an analogous result when θ is random and then making the prior converge to a degenerate distribution. The existence of the Bartlett correction to order $n^{-3/2}$ follows from the theorem. We show that an Edgeworth expansion with error $O(n^{-2})$ for T involves only polynomials of degree less than or equal to 3 and hence verify rigorously Lawley's (1956) result giving the order of the error in the Bartlett correction as $O(n^{-2})$.

1. Introduction. Let $X = (X_1, \dots, X_n)$ be a vector of observations with joint density $p(x,\theta)$, $\theta \in \Theta$ open $\subset R^p$, where we do not assume a priori any particular structure on $p(x,\theta)$. Consider the hypothesis H: $\theta^1 = \theta_0^1, \dots, \theta^k = \theta_0^k$. Suppose that maximum likelihood estimates $\hat{\theta}$ and $\hat{\theta}_H$ for $\theta \in \Theta$ and $\theta \in H$, respectively, are well defined. Then let

$$l(\theta) = n^{-1}\log p(X,\theta), \tag{1.1}$$

$$l(\hat{\theta}) = \max\nolimits_\Theta l(\theta), \tag{1.2}$$

$$l(\hat{\theta}_H) = \max\nolimits_H l(\theta) \tag{1.3}$$

and

$$\Lambda = 2n\left(l(\hat{\theta}) - l(\hat{\theta}_H)\right) \tag{1.4}$$

the usual likelihood ratio test statistic. All these quantities, of course, depend on n but we suppress this dependence to ease the notation. There is a common approximation to the distribution of Λ which has the status of a folk theorem:

$$L_\theta(\Lambda) \approx \chi_k^2$$

Received September 1987; revised July 1989.

[1]This paper was completed while the author was visiting AT & T Bell Telephone Labs, the Courant Institute and the University of Chicago.

[2]Research partially supported by ONR Contract N00014-80-C-0163.

AMS 1980 *subject classifications*. 62F05, 62F15.

Key words and phrases. Bartlett correction, signed log likelihood ratio statistic, Bernstein–von Mises theorem.

for $\theta \in H$. Theoretically this can be interpreted, for $\theta \in H$, as

$$P_\theta[\Lambda \le t] = \chi_k^2(t) + o(1) \tag{1.5}$$

as $n \to \infty$. This result was proved by Wilks (1938) and extended by Wald (1943) in the i.i.d. case, extended to the Markov case by Billingsley (1961) and subsequently extended to many other dependent and nonstationary situations. Bartlett (1937) noted, in the particular case of the hypothesis of the equality of variances for $k+1$ normal populations, that the χ_k^2 distribution was a far better fit to the distribution of $k\Lambda/E_\theta\Lambda$ than to Λ itself. Following work by Box (1949), Lawley (1956), by ingenious and difficult cumulant calculations, "established" the folk theorem that quite generally

$$P_\theta\left[\frac{k\Lambda}{\hat{E}} \le t\right] = \chi_k^2(t) + O(n^{-2}), \tag{1.6}$$

where

$$\hat{E} = k + \frac{\hat{b}}{n} = E_\theta(\Lambda) + O_p(n^{-3/2})$$

and $\hat{b}$ is a suitable estimate for the coefficient b of n^{-1} in the expansion of $E_\theta(\Lambda)$. Departing from an asymptotic formula for the conditional density of X given an ancillary due to Barndorff-Nielsen (1986). Barndorff-Nielsen and Cox (1984) showed that (1.6) can be expected to hold quite generally and they derived formulas for estimating b in one important class of models. Efron (1985) established (for an important special case) a related result. Let

$$T = \Lambda^{1/2}\operatorname{sgn}(\hat{\theta}^1 - \theta^1).$$

Then

$$P_\theta[T \le t] = \Phi\left(\frac{t - \mu(\theta)}{\sigma(\theta)}\right) + O(n^{-3/2}), \tag{1.7}$$

where

$$\mu(\theta) = \frac{a_0(\theta)}{\sqrt{n}} + \frac{a_1(\theta)}{n} + O(n^{-3/2}),$$

$$\sigma^2(\theta) = 1 + \frac{c(\theta)}{n} + O(n^{-3/2}),$$

where a_0, a_1 and c are suitable functions of θ, not depending on n. As P. McCullagh pointed out to us, this result implicitly already appears in Lawley (1956) and, in fact, $a_1 = 0$. It is easy to see that, for $k = 1$, (1.7) finally implies (1.6) [with $O(n^{-2})$ replaced by $O(n^{-3/2})$] with $\hat{b}$ estimating $a_0^2(\theta) + c(\theta)$.

Our aim in this paper is:

1. To give a generalization of Efron's result to vector parameters. A closely related result appears in Barndorff-Nielsen (1986) and is again foreshadowed by Lawley (1956).
2. To apply this extension to establish the validity of Bartlett's correction for the p variate joint distribution of the Λ statistics (deviances) arising from

testing the nested hypotheses H_k: $\theta^j = \theta_0^j$, $j = 1, \dots, k$, within H_{k-1} for $k = 1, \dots, p$. That is, to show that, when the deviances are standardized by their asymptotic expectations to order $1/n$, their joint distribution under θ_0 differs from that of p independent identically distributed χ_1^2 variables by an error of order n^{-2}. This result is also implicit in Lawley (1956) although the calculations are purely formal. For the case of a single statistic Λ, this can be obtained in a rigorous fashion under appropriate regularity conditions from Chandra and Ghosh (1979).
3. To give Bayesian analogues of both of these results which we believe provide a key to understanding the Bartlett phenomenon. The Bayesian analogue is interesting in its own right, is fairly easy to establish and is the basic step in our arguments for aims 1 and 2.

Here is a discussion of the motivation and the structure of our Bayesian argument when we restrict to the familiar case of i.i.d. observations from a smooth parametric family. It has been proved in Chandra and Ghosh (1979) that the distributions of the likelihood ratio, as well as Wald's and Rao's score statistic, have asymptotic expansions in powers of n^{-1}, which are valid in the sense of Bickel (1974). These types of expansions have been around for a long time; see Box (1949). When viewed as formal expansions for the density $p_n(\chi^2)$ of one of these statistics, they are of the form $ce^{-\chi^2/2}(\chi^2)^{k/2-1}\{1 + \psi_1(\chi^2)n^{-1} + \cdots\}$, where the coefficients ψ are polynomials in χ^2. It is easy to check that adjustment of such a statistic through multiplication or division by a constant of the form $(1 + bn^{-1})$ will knock off the coefficient of n^{-1} in the expansion for the adjusted statistic, iff ψ_1 is linear. By examining various examples one can convince oneself that ψ_1 is not linear for Wald's or Rao's statistic. Moreover it is far from clear why ψ_1 is linear for the likelihood ratio statistic. This paper is addressed to clearing up mysteries of this kind as well as to exploring the duality between the Bayesian and the frequentist setup which, to first order, was studied extensively by Le Cam under the rubric of the Bernstein–von Mises theorem.

Our Bayesian route could be followed to produce a relatively transparent proof of linearity of ψ_1. However, since we want to do more, namely, derive the asymptotic expansion for the joint distribution of the p deviances statistics up to $O(n^{-2})$, we first note, in a similar vein, that here also the question boils down to the structure of the polynomials that appear as coefficients of powers of n^{-1} in the expansion. The relevant results for this purpose are Lemmas A2 through A4 in the Appendix. These lemmas need to be applied to the vector $T(\theta, \mathbf{X})$ of the signed square roots of the likelihood ratio statistics, defined in Section 2. That the distribution of these statistics has a valid Edgeworth expansion can be shown using Theorem 2 of Bhattacharya and Ghosh (1978). In the frequentist setup the sort of structure one needs for the polynomials is specified in the conclusion of Theorem 3. It turns out that one needs the polynomials corresponding to $n^{-1/2}$ and n^{-1} to be of degree at most 1 and 2, respectively. To prove this, one first obtains a similar result in the Bayesian setup, namely, Theorem 1, which provides an expansion for the posterior for $\theta \in H$. Theoretically this can be interpreted, for $\theta \in H$, as

$$P_\theta[\Lambda \le t] = \chi_k^2(t) + o(1) \tag{1.5}$$

as $n \to \infty$. This result was proved by Wilks (1938) and extended by Wald (1943) in the i.i.d. case, extended to the Markov case by Billingsley (1961) and subsequently extended to many other dependent and nonstationary situations. Bartlett (1937) noted, in the particular case of the hypothesis of the equality of variances for $k + 1$ normal populations, that the χ_k^2 distribution was a far better fit to the distribution of $k\Lambda/E_\theta\Lambda$ than to Λ itself. Following work by Box (1949), Lawley (1956), by ingenious and difficult cumulant calculations, "established" the folk theorem that quite generally

$$P_\theta\left[\frac{k\Lambda}{\hat{E}} \le t\right] = \chi_k^2(t) + O(n^{-2}), \tag{1.6}$$

where

$$\hat{E} = k + \frac{\hat{b}}{n} = E_\theta(\Lambda) + O_p(n^{-3/2})$$

and $\hat{b}$ is a suitable estimate for the coefficient b of n^{-1} in the expansion of $E_\theta(\Lambda)$. Departing from an asymptotic formula for the conditional density of X given an ancillary due to Barndorff-Nielsen (1986). Barndorff-Nielsen and Cox (1984) showed that (1.6) can be expected to hold quite generally and they derived formulas for estimating b in one important class of models. Efron (1985) established (for an important special case) a related result. Let

$$T = \Lambda^{1/2}\,\mathrm{sgn}(\hat{\theta}^1 - \theta^1).$$

Then

$$P_\theta[T \le t] = \Phi\left(\frac{t - \mu(\theta)}{\sigma(\theta)}\right) + O(n^{-3/2}), \tag{1.7}$$

where

$$\mu(\theta) = \frac{a_0(\theta)}{\sqrt{n}} + \frac{a_1(\theta)}{n} + O(n^{-3/2}),$$

$$\sigma^2(\theta) = 1 + \frac{c(\theta)}{n} + O(n^{-3/2}),$$

where a_0, a_1 and c are suitable functions of θ, not depending on n. As P. McCullagh pointed out to us, this result implicitly already appears in Lawley (1956) and, in fact, $a_1 = 0$. It is easy to see that, for $k = 1$, (1.7) finally implies (1.6) [with $O(n^{-2})$ replaced by $O(n^{-3/2})$] with $\hat{b}$ estimating $a_0^2(\theta) + c(\theta)$.

Our aim in this paper is:

1. To give a generalization of Efron's result to vector parameters. A closely related result appears in Barndorff-Nielsen (1986) and is again foreshadowed by Lawley (1956).
2. To apply this extension to establish the validity of Bartlett's correction for the p variate joint distribution of the Λ statistics (deviances) arising from

testing the nested hypotheses H_k: $\theta^j = \theta_0^j$, $j = 1, \dots, k$, within H_{k-1} for $k = 1, \dots, p$. That is, to show that, when the deviances are standardized by their asymptotic expectations to order $1/n$, their joint distribution under θ_0 differs from that of p independent identically distributed χ_1^2 variables by an error of order n^{-2}. This result is also implicit in Lawley (1956) although the calculations are purely formal. For the case of a single statistic Λ, this can be obtained in a rigorous fashion under appropriate regularity conditions from Chandra and Ghosh (1979).

3. To give Bayesian analogues of both of these results which we believe provide a key to understanding the Bartlett phenomenon. The Bayesian analogue is interesting in its own right, is fairly easy to establish and is the basic step in our arguments for aims 1 and 2.

Here is a discussion of the motivation and the structure of our Bayesian argument when we restrict to the familiar case of i.i.d. observations from a smooth parametric family. It has been proved in Chandra and Ghosh (1979) that the distributions of the likelihood ratio, as well as Wald's and Rao's score statistic, have asymptotic expansions in powers of n^{-1}, which are valid in the sense of Bickel (1974). These types of expansions have been around for a long time; see Box (1949). When viewed as formal expansions for the density $p_n(\chi^2)$ of one of these statistics, they are of the form $ce^{-\chi^2/2}(\chi^2)^{k/2-1}\{1 + \psi_1(\chi^2)n^{-1} + \cdots\}$, where the coefficients ψ are polynomials in χ^2. It is easy to check that adjustment of such a statistic through multiplication or division by a constant of the form $(1 + bn^{-1})$ will knock off the coefficient of n^{-1} in the expansion for the adjusted statistic, iff ψ_1 is linear. By examining various examples one can convince oneself that ψ_1 is not linear for Wald's or Rao's statistic. Moreover it is far from clear why ψ_1 is linear for the likelihood ratio statistic. This paper is addressed to clearing up mysteries of this kind as well as to exploring the duality between the Bayesian and the frequentist setup which, to first order, was studied extensively by Le Cam under the rubric of the Bernstein–von Mises theorem.

Our Bayesian route could be followed to produce a relatively transparent proof of linearity of ψ_1. However, since we want to do more, namely, derive the asymptotic expansion for the joint distribution of the p deviances statistics up to $O(n^{-2})$, we first note, in a similar vein, that here also the question boils down to the structure of the polynomials that appear as coefficients of powers of n^{-1} in the expansion. The relevant results for this purpose are Lemmas A2 through A4 in the Appendix. These lemmas need to be applied to the vector $T(\theta, \mathbf{X})$ of the signed square roots of the likelihood ratio statistics, defined in Section 2. That the distribution of these statistics has a valid Edgeworth expansion can be shown using Theorem 2 of Bhattacharya and Ghosh (1978). In the frequentist setup the sort of structure one needs for the polynomials is specified in the conclusion of Theorem 3. It turns out that one needs the polynomials corresponding to $n^{-1/2}$ and n^{-1} to be of degree at most 1 and 2, respectively. To prove this, one first obtains a similar result in the Bayesian setup, namely, Theorem 1, which provides an expansion for the posterior distribution of $T(\theta, \mathbf{X})$ given $\mathbf{X}$. The likelihood factor in the posterior $\exp\{nl(\theta) - nl(\hat{\theta})\}$ is exactly the sum of squares of the components of T and so no expansion is needed. The coefficient polynomials in the asymptotic expansion arise only from the Taylor expansions of the prior density $\pi(\theta)$ around $\hat{\theta}$ and a stochastic expansion of the Jacobian of the transformation of $(\theta - \hat{\theta})$ to $T(\theta, X)$ viewed as a function of random θ. For reasons that are not hard to see, in these latter expansions the degree of the coefficient polynomial matches the power of n^{-1}; vide Lemmas 1 and 2. These facts are at the heart of the proof of Theorem 1. Theorem 1 would fail for Wald's or Rao's statistic because the likelihood factor $\exp\{nl(\theta) - nl(\hat{\theta})\}$ cannot be written as the square of either of them exactly and so an expansion of this term is called for too. Finally, Theorem 3 follows because Theorem 1 is true for a set of priors which is dense in the weak topology.

Our expansions may be used to set up Bayesian or frequentist confidence intervals; see the discussion following Corollary 1.

We propose to carry out our program without relying on the i.i.d. sampling assumption, under conditions such as those of Bickel, Götze and van Zwet (1985) which emphasize that we are, as with the original Wilks result, dealing with a phenomenon which depends only on the asymptotic stability of l and its derivatives, moderate deviation properties of $\hat{\theta}$ and related estimates and the existence of Edgeworth expansions for the distribution of T. Simple conditions implying those we give may be specified in the case of Markov and independent nonidentically distributed observations in the same way as is done in Bickel, Götze and van Zwet (1985).

A feature of our approach is that calculations are kept to a minimum so that, we believe, the phenomena are transparent. The disadvantage here is that unlike our predecessors, we do not arrive at formulae for the (estimated) coefficient $\hat{b}$ needed in the correction. It is, however, worth pointing out that, in situations which are like simple random sampling and where computing power is readily available, we can obtain $\hat{b}$ without knowing its form by applying the jackknife for bias reduction; see Efron (1982), for example. That is, we calculate Λ_{-i}, the Λ statistic for the data X_j, $j \neq i$, and put

$$\hat{b} = \sum_{i=1}^{n} (\Lambda_{-i}) - nk.$$

The paper is organized as follows. Section 2 contains the statements of the main theorems plus the necessary assumptions and notations. Section 3 contains the proofs of our results. Four simple technical lemmas are in the Appendix.

2. The main results. Since we intend to use tensor notation for arrays, we subsequently identify vector components by superscripts, for example, $\theta = (\theta^1, \dots, \theta^p)$. For given $\theta \in \Theta$, define $\hat{\theta}_j$ as the maximum likelihood estimate of θ when $\theta^1, \dots, \theta^j$ are fixed, i.e.,

$$l(\hat{\theta}_j) = \max\{l(\tau) : \tau^1 = \theta^1, \dots, \tau^j = \theta^j\}. \tag{2.1}$$

We shall in the sequel assume that these quantities exist and are unique but at the end of the section will sketch how this requirement can be weakened. Define $T = (T^1, \dots, T^p)$, where

$$T^j \equiv n^{1/2}\left[2\left(l\left(\hat{\theta}_{j-1}\right) - l\left(\hat{\theta}_j\right)\right)\right]^{1/2} \operatorname{sgn}\left(\hat{\theta}^j_{j-1} - \theta^j\right). \tag{2.2}$$

Note that T is a function of θ and $\mathbf{X}$.

Let π be a prior density on Θ. Let P denote the joint distribution of $(\theta, \mathbf{X})$ and $P(\cdot|\mathbf{X})$ the conditional (posterior) probability distribution of $(\theta, \mathbf{X})$ given $\mathbf{X}$. Let $r = n^{-1/2}$ and consider the posterior density of $r^{-1}(\theta - \hat{\theta})$ given by

$$\pi(h|\mathbf{x}) \equiv \exp\{l(\hat{\theta} + rh) - l(\hat{\theta})\}\pi(\hat{\theta} + rh)/N(\mathbf{x}),$$

where

$$N(\mathbf{X}) = \int \exp\{l(\hat{\theta} + rh) - l(\hat{\theta})\}\pi(\hat{\theta} + rh)\, dh. \tag{2.3}$$

Let

$$\phi(t) = (2\pi)^{-p/2} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{p} (t^i)^2\right\}$$

be the standard p variate normal density. Let $\pi_T(t|\mathbf{X})$ denote the posterior density of T [which exists under our assumptions with probability $1 - O(r^{m+1})$].

NOTATION. We postulate $m + 3$ continuous derivatives for $l(\theta), \pi(\theta)$ and write $l_{i_1 \cdots i_k}$ for $\partial^k l/\partial\theta^{i_1} \cdots \partial\theta^{i_k}$, etc. Following tensor notation, we indicate arrays by their elements. Thus l^i is a vector, l_{ij} a matrix, etc. We also follow the Einstein convention of summing over a subscript which is repeated in a superscript, e.g., $l_{ij}l^i = \sum_i l_{ij}l^i$. Occasionally we denote a vector array by symbols like ν_i, so that $\nu_i t^i$ stands for $\sum_i \nu_i t^i$.

Here are the main results stated under regularity conditions which appear at the end of the section.

THEOREM 1. *If B_m holds, then*

$$E_P \int |\pi_T(t|\mathbf{X}) - \pi_m(t, \mathbf{X})|\, dt = O(r^{m+1}), \tag{2.4}$$

where

$$\pi_m(t, \mathbf{X}) = \phi(t)(1 + P_m(r, \mathbf{X}, \pi) + Q_m(rt, \mathbf{X}, \pi))1(\mathbf{X} \in S),$$

P_m is a polynomial in r of degree m, Q_m is a polynomial in rt of degree m [both without constant terms and with coefficients which are rational functions of $l_{b_1 \cdots b_k}(\hat{\theta})$] and $\pi_{b_1 \cdots b_k}(\hat{\theta})/\pi(\hat{\theta})$ for $1 \le k \le n + 2$ and $P[X \notin S] = O(r^{m+1})$ where S is given in Section 3. 1(A), as usual, denotes the indicator of A.

Write

$$P_m(r, \mathbf{X}, \pi) = \sum_{k=1}^{m} P_{mk}(\mathbf{X}, \pi) r^k,$$

$$Q_m(u, \mathbf{X}, \pi) = \sum_{k=1}^{m} Q_{mb_1 \cdots b_k}(\mathbf{X}, \pi) u^{b_1} \cdots u^{b_k}$$

and note that P_m, Q_m and S depend on n.

NOTE 1. It is necessary to keep the indicator of S in π_m since the coefficients $P_{mk}, Q_{mb_1 \cdots b_k}$ need not be bounded outside S.

The proof of Theorem 1 actually also yields that if $X \in S$, i.e., with probability $1 - O(r^{m+1})$, the random quantity

$$\int |\pi_T(t|\mathbf{X}) - \pi_m(t|\mathbf{X})| \, dt$$

is $O(r^{m+1})$.

NOTE 2. Since P_{mk} and $Q_{mb_1 \cdots b_k}$ depend on r they are not uniquely defined. Since

$$E\left|\int \pi_m(t, \mathbf{X}) \, dt - 1\right| = O(r^{m+1}). \tag{2.4'}$$

It is easy to see that we can always take $P_{m0} = Q_{m0} = 0$ and suppose all P_{mk} for k odd to be zero. For example, suppose we are given a set of $P_{mk}^{(1)}$ and associated Q_m. Note that

$$P_{m1}^{(1)} r + \int Q_{m1} r t \phi(t) \, dt = O(r^2) \quad \text{if } m \geq 1.$$

Therefore, $P_{m1}^{(1)} = O(r)$. Hence we can define the following set $P_{mk}^{(2)}$ satisfying (2.4): $P_{m0}^{(2)} = 0$, $P_{m2}^{(2)} = P_{m1}^{(1)} r + P_{m2}^{(1)} + P_{m3}^{(1)} r$, $P_{mk}^{(2)} = 0$ for k odd and $P_{mk}^{(2)} = P_{mk}^{(1)} + r P_{m(k+1)}^{(1)}$ for k even and greater than or equal to 4.

NOTE 3. Note that (2.4') for $m = 2, 3$ implies

$$E\left|\int \pi_2(t, X) \, dt - 1\right| = E\left|P_{22} r^2 - Q_{2ij} \delta^{ij} r^2\right| = O(r^3).$$

In view of Notes 1 and 2 and the above relation we deduce, putting $m = 1, 2$ in (2.4), that with probability $1 - O(r^2)$ and $1 - O(r^3)$, respectively, the posterior distribution of T is $N_p(rQ_{1i}, J)$ with error $O(r^2)$ and $N_p(rQ_{2i}, J + r^2(2Q_{2ij} - Q_{2i}Q_{2j}))$ with error $O(r^3)$, where $N_p(\mu, \Sigma)$ is the p variate normal distribution with mean μ and dispersion matrix Σ and J is the $p \times p$ identity matrix. These are the multivariate Bayesian analogues of Efron's (1985) result.

NOTE 4. The relation (2.4') for $m = 3$ implies as above that

$$E\left|P_{32} r^2 - Q_{3ij} \delta^{ij} r^2\right| = O(r^4)$$

and hence that π_3 may be written as

$$rQ_{3i}t^i + r^2Q_{3ij}(t^it^j - \delta^{ij}) + r^3Q_{3ijk}t^it^jt^k + O(r^4),$$

which has the structure of $g(t)$ of Lemma A2 up to $O(r^4)$. This fact will be used in the proof of Theorem 2.

Let $c_k(\cdot)$ denote the χ_k^2 density,

$$D^j \equiv (T^j)^2 = 2n\left(l\left(\hat{\theta}_{j-1}\right) - l\left(\hat{\theta}_j\right)\right)$$

the deviance and

$$\tilde{D}^j = D^j/\left(1 + 2r^2Q_{2jj}\right)$$

the standardized (Bartlett corrected) deviance. If π_D and $\pi_{\tilde{D}}$ are the corresponding posterior densities of these vectors $D = (D^1, \dots, D^p)$ and $\tilde{D} = (\tilde{D}^1, \dots, \tilde{D}^p)$, then one has the following result.

THEOREM 2. *Under* B_1,

$$E_P\left\{\int\left|\pi_D(u|\mathbf{X}) - \prod_{j=1}^p c_1(u^j)\right| dt\, 1(\mathbf{X} \in S)\right\} = O(n^{-1}), \tag{2.5}$$

while under B_3,

$$E_P\left\{\int\left|\pi_{\tilde{D}}(u|\mathbf{X}) - \prod_{j=1}^p c_1(u^j)\right| du\, 1(\mathbf{X} \in S)\right\} = O(n^{-2}). \tag{2.6}$$

In fact (vide Note 1), with probability $1 - O(n^{-1})$ and error $O(n^{-1})$ the posterior distribution of D is that of p independent χ_1^2, while for $\tilde{D}$ the same claim holds with probability $1 - O(n^{-2})$ and error $O(n^{-2})$.

From this we deduce:

COROLLARY 1. (a) *Under* B_1, *if* π_Λ *is the posterior distribution of* Λ *given by* (1.4),

$$E_P\left\{\int|\pi_\Lambda(u|\mathbf{X}) - c_k(u)|\, du\, 1(\mathbf{X} \in S)\right\} = O(n^{-1}). \tag{2.7}$$

(b) *Let* $\tilde{\Lambda} = \Lambda/(1 + 2r^2k^{-1}\Sigma_{j=1}^k Q_{2jj})$. *Then, under* B_3,

$$E_P\left\{\int|\pi_{\tilde{\Lambda}}(u|\mathbf{X}) - c_k(u)|\, du\, 1(\mathbf{X} \in S)\right\} = O(n^{-2}). \tag{2.8}$$

So (2.7) says that the posterior distribution of Λ is χ_k^2 with error $O(n^{-1})$ while (2.8) is the Bayesian analogue of the Bartlett phenomenon. The posterior distribution of the Bartlett standardized statistic $\tilde{\Lambda}$ is χ_k^2 with error $O(n^{-2})$.

These results can in principle be used to set Bayesian posterior confidence regions for θ to order n^{-1}, n^{-2} in a variety of ways. For instance, $\{\theta\colon \Lambda \le \chi_p(1 - \alpha)\}$ where χ_p is the $1 - \alpha$ percentile of χ_p^2 and $\Lambda = 2(l(\hat{\theta}) - l(\theta))$ has posterior probability $1 - \alpha$ with error $O(n^{-1})$, while $\{\theta\colon \tilde{\Lambda} \le \chi_p(1 - \alpha)\}$

has posterior probability $1 - \alpha$ with error $O(n^{-2})$. Of course regions could be based on other functions of D_j and $\tilde{D}_j$, for instance, on $\max_j D_j$ or $\max_j \tilde{D}_j$. They could also be used in investigating the old question of what choices of model and prior lead to posterior probability regions which are also frequentist regions with error $O(n^{-2})$; see, for example, Stein (1985) and Welch and Peers (1963). However, more detailed computation of the Q_j than we provide seems necessary for this endeavor.

We use these results only in establishing the corresponding result in the frequentist case.

THEOREM 3. *Suppose that F_m holds and the density of T, $p_T(t|\theta)$, admits an Edgeworth expansion such that if $i^2 = -1$,*

$$(2.9) \qquad \left| \int e^{i\nu_j t^j} \left[p_T(t|\theta) - \phi(t)\left\{1 + \sum_{k=1}^{m} r^k R_k(t,\theta)\right\}\right] dt \right| = O(r^{m+1})$$

uniformly in compact sets of θ and ν, where the $R_k(\cdot,\theta)$ are continuous in θ and polynomials in t, independent of r. Then, the R_k are of at most degree k in t.

As in Notes 2 and 3, it is clear that (2.9) implies, on taking $\nu = 0$, that $R_1(t,\theta) = R_{1j}t^j$ and $R_2(t,\theta) = R_{2ij}(t^i t^j - \delta^{ij}) + R_{2i}t^i$, where δ^{ij} is the Kronecker delta. In the following we shall need a condition analogous to (2.9), namely,

$$(2.9') \qquad \left| \int e^{i\nu_j (t^j)^2} \left[p_T(t|\theta) - \phi(t)\left\{1 + \sum_{k=1}^{m} r^k R_k(t,\theta)\right\}\right] dt \right| = O(r^{m+1})$$

uniformly in compact sets of θ and all ν. We deduce our generalization of Efron's result.

COROLLARY 2. *If $m = 1$, the characteristic function of p_T differs from that of $N(rR_{1j}, J)$ by $O(r^2)$ and if $m = 2$, from $\mathbf{N}(rR_{ij}, J + r^2(2R_{2j} - R_{1i}R_{1j}))$ by $O(r^3)$.*

THEOREM 4. *If the assumptions of Theorem 3 and (2.9′) hold for $m = 1$, then, uniformly in ν,*

$$(2.10) \qquad \int e^{i\nu u}\left[p_D(u|\theta) - \prod_{j=1}^{p} c_1(u^j)\right] du = O(n^{-1}),$$

i.e., the approximation $\prod_{j=1}^{p} c_1(u^j)$ is good to order n^{-1}.

Further, let

$$\tilde{D}^j = D^j / \left(1 + 2r^2 R_{2jj}\right).$$

If (2.9), (2.9′) and F_m hold for $m = 3$, then uniformly in ν,

$$(2.11) \qquad \int e^{i\nu_j u^j}\left[p_{\tilde{D}}(u|\theta) - \prod_{j=1}^{p} c_1(u^j)\right] du = O(n^{-2}).$$

COROLLARY 3. *Under the conditions of Theorem* 4, *uniformly in* ν,

$$\int e^{i\nu u}[p_{\Lambda}(u|\theta) - c_k(u)]\,du = O(n^{-1}), \tag{2.12}$$

$$\int e^{i\nu u}[p_{\tilde{\Lambda}}(u|\theta) - c_k(u)]\,du = O(n^{-2}). \tag{2.13}$$

It turns out that $T^i = r^{-1}(\hat{\eta}^i - \eta^i) + O(r)$ [see (3.6) and (3.19)] and $r^{-1}(\hat{\eta}^i - \eta^i)$ is up to $O(r)$ a linear function of the first derivatives of the log likelihood evaluated at θ. In fact it is possible to stochastically expand T in terms of the derivatives of the log likelihood evaluated at θ, with a leading linear term. In the i.i.d. case if enough moments are finite, we can talk of a formal Edgeworth expansion for the density or distribution function of T and under the same assumptions the rigorous expansion of the characteristic function of T that we require is valid; vide the introduction in Bhattacharya and Ghosh (1978). This is all that one needs to justify the Bartlett correction and the related results as given in Theorem 4. If one wants these results to be valid for the distribution function in the sense of Bickel (1974), it is enough to assume that the Edgeworth expansion for the density of T is valid in the L_1 sense. This assumption may be verified via Theorem 2(a) of Bhattacharya and Ghosh (1978) if the derivatives of the log likelihood appearing in the stochastic expansion for T up to $o_p(n^{-3/2})$ have an absolutely continuous joint distribution. Actually, instead of absolute continuity, it is enough to assume Cramer's condition [vide condition C of Bhattacharya and Ghosh (1978)] and apply their Theorem 2(b) instead of Theorem 2(a).

We note again that a form of Theorem 4 appeared in Barndorff-Nielsen (1986) [with error $O(n^{-3/2})$]. Barndorff-Nielsen's results focus on conditional inference given asymptotic ancillary statistics. His work implicitly requires conditions for the validity of saddlepoint expansions for the conditional density. These in turn imply but are not necessary for the validity of Edgeworth expansions for the conditional density. The Edgeworth expansions may be used in conjunction with our "Bayesian" result to derive the appropriate analogues of Theorem 4. We believe our Bayesian route makes matters easier and more transparent. The assumptions below may appear rather strong but, as indicated in the remarks, they hold quite generally. Moreover, they are quite natural if one is to develop a rigorous, rather than a formal, argument.

Suppose we estimate the correction factor and adjust the likelihood ratio statistic in (1.6). If in Corollary 3 we replace $\tilde{\Lambda}$ by $k\Lambda/(k + \hat{b}/n)$ then the conclusion of Corollary 3 holds under suitable regularity conditions. This fact was first noted by Barndorff-Nielsen and Hall (1988). The most brutal condition is to suppose that

$$\hat{b} = b(\theta) + rc_i t^i + \Delta(\theta), \tag{2.14}$$

where

$$E_\theta|\Delta(\theta)| = O(r^2).$$

Of course (2.14) is motivated by a stochastic expansion such as

$$\hat{b} \equiv b(\hat{\theta}) = b(\theta) + d_i(\hat{\theta}^i - \theta^i) + O_p(r^2) \tag{2.15}$$

and the expansion

$$\hat{\theta}^i - \theta^i = r\hat{D}_{ij}T^i + O_p(r^2)$$

for a suitable $\hat{D}_{ij}$; see Lemma 2. To show that (2.14) and the assumptions of Corollary 3 are enough for this result we need only note that the difference between the Fourier transforms of $\tilde{\Lambda}$ and $k\Lambda/(k + \hat{b}/n)$ at ν can be written [with an appropriate constant $M(\theta)$] as

$$M(\theta)\int \exp\left[\left(-\tfrac{1}{2}\sum_{i=1}^{p}[t^i]^2\right) + i\nu\Sigma[t^i]^2\right]\left[\Sigma[t^i]^2(c_i t^i)\right]r^3\,dt + O(r^4)$$

uniformly on compact ν subsets. The integral vanishes by symmetry.

Condition (2.14) is too brutal but can readily be replaced by the possibility of further expansion of (2.15) and large deviation estimates for $\hat{\theta} - \theta$. Alternatively, we can simply suppose that the Edgeworth expansion of $k\Lambda(k + \hat{b}r^2)^{-1}$ agrees with that of $\tilde{\Lambda}(1 - (k + b(\theta)r^2)^{-1}r^2c_iT^i)$ with error of order r^2. This kind of replacement can be proved in a standard fashion under the usual protocols for asymptotic expansions of maximum likelihood estimates; see Pfanzagl (1974), for example.

We postulate nonrandom arrays λ_i, λ_{ij}, etc. and write,

$$l_{i_1\cdots i_k}(\theta) = \lambda_{i_1\cdots i_k}(\theta) + \Delta_{i_1\cdots i_k}(\theta).$$

Here are our conditions. Let $|\cdot|$ denote the l_1 norm on R^p. For all $0 < M < \infty$ and some $0 < \delta < 1$, $\varepsilon_n \downarrow 0$.

B_m: (i) $P[|\hat{\theta} - \theta| \geq Mr^{1-\delta}] = O(r^{m+1})$.

(ii) $P[|\hat{\theta} - \theta| \leq Mr^{m+2}] = O(r^{m+1})$.

Let

$$A = \left\{\mathbf{x}\text{: for all } j, \left\{\theta: |\hat{\theta}(\mathbf{x}) - \theta| \leq M_1 r^{1-\delta}\right\} \subset \left\{\theta: |\hat{\theta}_j(\mathbf{x},\theta) - \hat{\theta}(\mathbf{x})| \leq M_2 r^{1-\delta}\right\}\right\}.$$

For all $0 < M_1 < \infty$, there exists $0 < M_2 < \infty$ such that:

(iii) $P[\mathbf{X} \notin A] = O(r^{m+1})$.

(iv) $P[\sup\{|\Delta_{i_1\cdots i_k}(\hat{\theta} + r\nu)|: |\nu| \leq Mr^{1-\delta}\} \geq \varepsilon_n] = O(r^{m+1})$, $1 \leq k \leq m + 3$.

(v) The maps $\theta \to \lambda_{i_1\cdots i_k}(\theta)$ are continuous, $1 \leq k \leq m$.

(vi) The matrix $\|-\lambda_{ij}(\theta)\|$ is positive definite for all θ.

(vii) (a) π vanishes off a compact $K \subset \Theta$. (b) $P[\sup\{|\pi_{i_1\cdots i_{m+2}}(\hat{\theta} + r\nu)|/\pi(\hat{\theta})$: $|\nu| \leq Mr^{-\delta}\} \geq r^{-\delta}] = O(r^{m+1})$.

F_m: Uniformly on compacts in θ:

(i) $P_\theta[|\hat{\theta} - \theta| \geq Mr^{1-\delta}] = O(r^{m+1})$.

(ii) $P_\theta[|\hat{\theta} - \theta| \leq Mr^{m+2}] = O(r^{m+1})$.

(iii) $P_\theta[\mathbf{X} \notin A] = O(r^{m+1})$ for A defined in B_m.

(iv) $P_\theta[\sup\{|\Delta_{i_1\cdots i_k}(\theta + r\nu)|: |\nu| \leq Mr^{1-\delta}\} \geq \varepsilon_n] = O(r^{m+1})$, for $1 \leq k \leq m + 3$.

(v) Condition (v) of B_m.

(vi) Condition (vi) of B_m.

REMARKS. (a) We give a qualitative discussion of the "Bayesian" conditions B_m. The frequentist conditions F_m can be viewed in an analogous fashion.

(i) Variations of the mle $\hat{\theta}$ from θ of order $n^{-1/2(1-\delta)}$ occur with very small probability. Thus we can safely think about Taylor expanding $l(\theta)$ and $l(\hat{\theta}_j(\theta))$ around $\hat{\theta}$.

(ii) This condition says that $r^{-1}(\hat{\theta} - \theta)$ has approximately a bounded density near 0. It is needed to ensure that the map $\theta - \hat{\theta} \to \mathbf{T}(\theta, \mathbf{x})$ is 1–1 and otherwise well behaved with high probability.

(iii) This condition assumes that both $\hat{\theta}$ and $\hat{\theta}_j$ are close to θ and each other

simultaneously. It is needed for expansions of $l(\hat{\theta}_j(\theta))$.

(iv) The coefficients of the Taylor expansion differ little from constants, or more specifically, $l(\theta)$ and its derivatives behave like averages of i.i.d. variables.

(v) Smoothness conditions needed to permit replacement of quantities such as $\lambda_{i_1 \cdots i_k}(\hat{\theta}_j(\theta))$ appearing as approximations to coefficients in the Taylor expan-

sion of $l(\hat{\theta}_j(\theta))$ by $\lambda_{i_1 \cdots i_k}(\hat{\theta})$.

(vi) Nonsingularity of the information matrix is necessary even for the statement of the Bernstein–von Mises theorem.

(vii) We need to expand $\log \pi(\theta)$ around $\hat{\theta}$. Condition (a) is useful for technical reasons, while (b) is needed to control $\log \pi$ and its derivatives near the boundary of K where $\log \pi \to -\infty$.

(b) The validity of F_m and B_m other than (ii) and (iii) has been checked for independent nonidentically distributed and Markov dependent observations in Bickel, Götze and van Zwet (1985). In particular these conditions hold for exponential families in the i.i.d. case. They also hold in many examples for such families in the independent nonidentically distributed case, e.g., in regression and GLIM models. Another example is the class of aperiodic irreducible finite state Markov chains with stationary completely unknown transition matrix.

(c) Condition B_m(ii) in fact follows from the other B_m conditions since they guarantee an Edgeworth expansion for $\pi(h|\mathbf{X})$. An Edgeworth expansion uniform on θ compacts for the distribution of $r^{-1}(\hat{\theta} - \theta)$ implies F_m(i) and (ii). Condition F_m or B_m(iii) holds if the log likelihood is convex.

(d) The conditions on existence of the estimate $\hat{\theta}_j$ can be replaced by requiring the existence of a preliminary estimate $\tilde{\theta}$ with appropriate moderate deviation properties and then redefining the $\hat{\theta}_j$ as the result of $m + 1$ iterations of the Newton–Raphson method applied to the appropriate likelihood equations. See Theorem 4 of Bickel, Götze and van Zwet (1985).

(e) In the situation of (d), suppose that F_m(iv)–(vi) hold and that, uniformly on θ compacts, for all $0 < M < \infty$,

$$P_\theta\big[|\tilde{\theta} - \theta| \geq M r^{1-\delta}\big] = O(r^{m+1}),$$
$$P_\theta\big[|\tilde{\theta} - \theta| \leq M r^{m+2}\big] = O(r^{m+1}). \tag{2.16}$$

Let

$$A^* = \left\{\mathbf{x}\colon \text{for all } j, \{\theta\colon |\tilde{\theta} - \theta| < M_1 r^{1-\delta}\} \subset \left\{\theta\colon \left|\hat{\theta}_j - \tilde{\theta}\right| < M_2 r^{1-\delta}\right\}\right\}.$$

Then uniformly on θ compacts,

$$P_\theta[\mathbf{X} \in A^*] = O(r^{m+1}).$$

If we redefine the set B of Section 3 so that B(ii) is replaced by

$$|\hat{\theta}^b - \theta^b| > M^* r^{m+2}, \qquad |\tilde{\theta}^b - \theta^b| < r^{1-\delta},$$

then the proof of Theorems 4 and 5 goes through.

3. Proofs. We need to analyze $\pi_T(t|\mathbf{X})$ where we assume that $\mathbf{X}$ belongs to
a set S on which the map $h \to T(\hat{\theta} + rh, \mathbf{X})$, $|h| < Mr^{-\delta}$, is invertible with nonvanishing Jacobian and the matrix $\| -l_{ij}(\hat{\theta})\| = \hat{C}$ is positive definite. We explain the transformation in more detail and give S below. Let $\hat{D}$ be the unique lower triangular matrix with positive diagonal such that

$$\hat{D}\hat{D}^T = \hat{C} \tag{3.1}$$

and

$$L(\eta) = l(\hat{D}^{-1}\eta). \tag{3.2}$$

If $\|l_{ij}(\hat{\theta})\|$ is the Hessian of l at $\hat{\theta}$ and $\hat{\eta} = \hat{D}\hat{\theta}$, then in the usual notation,

$$-L_{ij}(\hat{\eta}) = J, \tag{3.3}$$

the $p \times p$ identity. This in the Bayesian domain corresponds to standardizing the Fisher information at θ to be J as is done in the corresponding frequentist calculations. Further define $\hat{\eta}_j$ by

$$L(\hat{\eta}_j) = \max\{L(\gamma)\colon \gamma^1 = \eta^1, \dots, \gamma^j = \eta^j\} \tag{3.4}$$

and

$$\tilde{T}^i(\eta) = r(2(L(\hat{\eta}_{i-1}) - L(\hat{\eta}_i)))^{1/2} \operatorname{sgn}(\hat{\eta}^i_{i-1} - \eta^i). \tag{3.5}$$

It is easy to verify that

$$T(\hat{\theta} + rh) = \tilde{T}(\hat{\eta} + r\hat{D}h). \tag{3.6}$$

Now $\hat{D}r^{-1}(\theta - \hat{\theta})$ has posterior density

$$\pi(\hat{D}^{-1}h|\mathbf{X})\left|\det(\hat{D})\right|^{-1} \tag{3.7}$$

and hence

$$\pi_T(t|\mathbf{X}) = \exp\left(-\tfrac{1}{2}\sum_{i=1}^{p}(t^i)^2\right)\pi\left(\hat{D}^{-1}(\hat{\eta} + rh(t))\right)\det\left\|h^i_j(t)\right\|/M(\mathbf{X}), \tag{3.8}$$

where $h(t)$ is defined by

$$\tilde{T}(\hat{\eta} + rh(t)) = t \tag{3.9}$$

and

$$h_j^i(t) = \frac{\partial h^i}{\partial t_j}(t),$$

$$M(\mathbf{X}) = \int \exp\left(-\tfrac{1}{2}\sum_{i=1}^{p}(t^i)^2\right)\pi\big(\hat{D}^{-1}(\hat{\eta} + rh(t))\big)\det\big\|h_j^i(t)\big\|.$$

For fixed $\mathbf{X}$, let R_X be the image of $\{h\colon |h| < Mr^{-\delta}\}$ under the map $h \to T(\hat{\theta} + rh, \mathbf{X})$. From (3.8) it is clear that our task in proving Theorem 1 is to exhibit the set S such that, for $t \in R_X$, h is uniquely defined by (3.9) and such that

$$h(t) = t + rP(t,\mathbf{X}) + O(r^{m+1}), \tag{3.10}$$

$$h_j^i(t) = \delta_{ij} + rP_{ij}(t,\mathbf{X}) + O(r^{m+1}), \tag{3.11}$$

where P and P_{ij} are polynomials in t, and to identify the order of the polynomials. Here $O(r^{m+1})$ means that the remainder is bounded on S by Mr^{m+1} for a generic constant M independent of n.

We define B as the set where

(i) $\sup\{|\pi_{i_1 \cdots i_{m+2}}(\hat{\theta} + r\nu)|/\pi(\hat{\theta})\colon |\nu| \le Mr^{-\delta}\} \le r^{-\delta}$.

(ii) $M^* r^{m+2} < |\hat{\theta}^b - \theta^b| < r^{1-\delta}, \quad 1 \le b \le p$.

(iii) $\sup\{|\Delta_{i_1 \cdots i_k}(\hat{\theta} + r\nu)|\colon |\nu| \le Mr^{-\delta}\} \le \varepsilon_n$.

Note that, by B_m,

(a) $P[(r^{-1}(\hat{\theta} - \theta), \mathbf{X}) \in B^c] = O(r^{m+1})$.

(b) The $\mathbf{x}$ sections of B intersect each quadrant in an open convex set since $|\cdot|$ is the l_1 norm.

(c) There exists a generic constant $C > 0$ such that on B,

$$\sup\left\{\left|l_{i_1 \cdots i_k}(\hat{\theta} + rh)\right|\colon |h| \le Mr^{-\delta}\right\} \le C.$$

(d) $C^{-1} \le \lambda \le \bar{\lambda} \le C$ where $\lambda, \bar{\lambda}$ are the minimal and maximal eigenvalues of $\| -l_{ij}(\hat{\theta})\|$.

(e) $|\hat{\theta}_j - \hat{\theta}_{j-1}| \le M_2 r^{1-\delta}, |\hat{\theta}_j - \hat{\theta}| \le M_1 r^{1-\delta}$.

We let $\tilde{S}$ be the image of B under the map $(h, x) \to (T(\theta(x) + rh, x), x)$ and S be just the projection of $\tilde{S}$ on the $\mathbf{x}$ axis, i.e., the set of all $\mathbf{x}$ satisfying (i) and (iii) above.

CONVENTION. Expressions such as $\hat{\eta}_i(\eta)$ are calculated at $\eta = \hat{\eta} + rh$.

LEMMA 1. *On B, for $j \ge i + 1$,*

$$\hat{\eta}_i^j = \hat{\eta}^j + \sum_{k=2}^{m+1} N_{b_1 \cdots b_k}^{ij} r^k h^{b_1} \cdots h^{b_k} + O(r^{m+1}), \tag{3.12}$$

where $N_{i_1 \cdots i_k}$ are polynomials in the derivatives $L_{i_1 \cdots i_t}$ of L (evaluated at $\hat{\eta}$)

with $t \le k$ *and* $h = r^{-1}(\eta - \hat{\eta})$ *with no constant term. Let* $d = \hat{\eta}^i_{i-1} - \eta^i$. *Then*

$$\hat{\eta}^j_{i-1} - \hat{\eta}^j_i = \sum_{k=1}^{m+1} M^{ij}_k d^k + O(|d|^{m+2}), \tag{3.13}$$

where M^{ij}_k *are polynomials in* $L_{i_1 \cdots i_k}$ *and* rh *which vanish at* $h = 0$.

PROOF. Write L_{ab}, etc., for derivatives of L evaluated at $\hat{\eta}$. For $j \ge i + 1$,

$$\begin{aligned} 0 = L_j(\hat{\eta}_i) - L_j(\hat{\eta}) &= L_{jb}(\hat{\eta}^b_i - \hat{\eta}^b) + \cdots + \frac{1}{(m+1)!} L_{j,b_1 \cdots b_{m+1}} \\ &\quad \times \prod_{k=1}^{m+1} (\hat{\eta}^{b_k}_i - \hat{\eta}^{b_k}) + O(r^{m+1}). \end{aligned} \tag{3.14}$$

To see this, note first that $\hat{\eta}_i = \hat{D}\hat{\theta}_i$ and hence, in view of (e), $|\hat{\eta}_i - \hat{\eta}| \le M_3 r^{1-\delta}$. Therefore, applying (c) and (d), again the relevant derivatives of order up to $m + 2$ of L at $\hat{\eta}$ are bounded and (3.14) follows. Note that by (3.3), $L_{ab} = -\delta_{ab}$ and that

$$\hat{\eta}^b_i - \hat{\eta}^b = -rh^b \quad \text{for } b \le i.$$

So we can rewrite (3.14) in the form

$$\delta_{jb} u^b = P_j(u, rh) + O(r^{m+1}), \qquad j \ge i + 1, \tag{3.15}$$

where $u^b = \hat{\eta}^b_{i-1} - \hat{\eta}^b_i$ and P_j is a polynomial of degree $(m + 1)$ in u and rh with no term of combined degree less than 2 and bounded coefficients which are polynomials in the $L_{i_1 \cdots i_t}$.

Claim (3.12) follows from a standard Lagrange inversion argument. For (3.13) write, for $j \ge i + 1$,

$$0 = L_j(\hat{\eta}_i) - L_j(\hat{\eta}_{i-1}) = -L_{jb}(\hat{\eta}^*) e^b, \tag{3.16}$$

where $\hat{\eta}^*$ is an intermediate value and $e^b = \hat{\eta}^b_{i-1} - \hat{\eta}^b_i$.

Note that

$$e^b = 0, \qquad b \le i - 1, \qquad e^i = d \tag{3.17}$$

and

$$L_{jb}(\hat{\eta}^*) = -\delta_{jb} + O(r),$$

so that (3.16) yields, for $j \ge i + 1$,

$$|\hat{\eta}^j_{i-1} - \hat{\eta}^j| = O(r)|d|. \tag{3.18}$$

Expand further to get

$$\begin{aligned} &L_{jb}(\hat{\eta}_{i-1}) e^b + \cdots + \frac{1}{(m+1)!} L_{jb_1 \cdots b_{m+1}}(\hat{\eta}_{i-1}) e^{b_1} \cdots e^{b_m} \\ &\quad + O(|d|^{m+2}) = 0. \end{aligned} \tag{3.19}$$

Rewrite (3.19) in the form

$$A_{jb}e^{b} + A_{jb_1b_2}e^{b_1}e^{b_2} + A_{jb_1\cdots b_{m+1}}e^{b_1}\cdots e^{b_{m+1}}$$
$$= a_1 d + \cdots + a_{m+1}d^{m+1} + O(d^{m+1}),$$

where the indices $b, b_1, \dots, b_m$ range from $i+1$ to p,

$$A_{jb_1\cdots b_k} = \frac{L_{jb_1\cdots b_k}}{k!}(\hat{\eta}_{i-1})$$

and the a_i are polynomials in the $L_{jb_1\cdots b_k}(\hat{\eta}_{i-1})$ and the e^b. Expand $A_{jb_1\cdots b_k}$ around $\hat{\eta}$ to $m+1-k$ terms and use (3.12) to conclude that with remainder $O(r^{m+1})$, all the $A_{jb_1\cdots b_k}$ are polynomials in $L_{jb_1\cdots b_t}$ and rh. Finally note that, for $b \geq i+1$,

$$e^b = \hat{\eta}^b_{i-1} - \hat{\eta}^b_i = (\hat{\eta}^b_{i-1} - \hat{\eta}^b) - (\hat{\eta}^b_i - \hat{\eta}^b)$$

can by (3.12) itself be written as a polynomial of rh and $L_{jb_1\cdots b_t}$ so that the a_j are also, up to order $m+1$, polynomials in rh and $L_{jb_1\cdots b_t}$, for $t \leq m+1$. The lemma follows. □

LEMMA 2. *On B*

$$\tilde{T}^i(\hat{\eta} + rh) = h^i + r^{-1}Q^i(rh) + O(r^{m+1}),$$

where Q is a polynomial of degree $m+1$ in rh with no constant or linear term and coefficients which are polynomials in $L_{b_1\cdots b_k}$, $k \leq m+2$.

PROOF. By definition

$$\tilde{T}^i(\hat{\eta} + rh) = r^{-1}\Bigg[-\sum_{k=1}^{m+2}\frac{2}{k!}L_{b_1\cdots b_k}(\hat{\eta}_{i-1})\prod_{t=1}^{k}(\hat{\eta}^{b_t}_{i-1} - \hat{\eta}^{b_t}_i) \tag{3.20}$$
$$+O(|\hat{\eta}_{i-1} - \hat{\eta}_i|^{m+3})\Bigg]^{1/2}\operatorname{sgn}(\hat{\eta}^i_{i-1} - \eta^i).$$

Note that $L_b(\hat{\eta}_{i-1}) = 0$, $b \geq i$, and $\hat{\eta}^b_{i-1} = \hat{\eta}^b_i$, $b \leq i-1$, so that the first term in the sum vanishes. Expand the coefficients around $\hat{\eta}$ and use (3.18) and (3.13) to get

$$\tilde{T}^i(\hat{\eta} + rh) = r^{-1}\left(d + \sum_{k=2}^{m+2}c_k d^k + O(r^{-1}|d|^{m+2})\right), \tag{3.21}$$

where the c_k are polynomials in rh. Now substitute for d from (3.12),

$$d = rh^i + \sum_{k=2}^{m+1}N^{i-1,i}_{b_1\cdots b_k}r^k h^{b_1}\cdots h^{b_k} + O(r^{m+1}), \tag{3.22}$$

and the lemma follows. □

LEMMA 3. (i) *If O_i, $i = 1, \dots, 2^p$, are the quadrants of R^p, then $\tilde{T}(\hat{\eta} + rh)$ maps $O_i \cap B_{\mathbf{x}}$ into O_i for all i.*

(ii) *$\tilde{T}$ is continuously differentiable on $O_k \cap B_{\mathbf{x}}$ for $1 \leq k \leq 2^p$. Let*

$$\tilde{T}^i_j = \frac{\partial \tilde{T}^i}{\partial h_j}.$$

Then $\tilde{T}^i_j$ *is lower triangular and*

$$\tilde{T}^i_i = 1 + P^i(rh) + O(r^{m+1}), \tag{3.23}$$

where P^i *is a polynomial of degree* $m + 1$ *with no constant term and coefficients in* $L_{b_1 \cdots b_k}$, $k \le m + 2$.

(iii) $\tilde{T}$ *is* 1–1.

PROOF. (i) We need to show that on B,

$$\operatorname{sgn}(\hat{\eta}^i_{i-1} - \eta^i) = \operatorname{sgn} h, \qquad i = 1, \dots, p. \tag{3.24}$$

By (3.12) on B,

$$\hat{\eta}^i_{i-1} - \eta^i = rh^i(1 + rM_1(h)) + r^{m+2}M_2(h),$$

where M_1 is a polynomial in h with bounded coefficients and $|M_2(h)|$ is bounded by M_2 for all $(\mathbf{x}, h) \in B$. But $(\mathbf{x}, h) \in B \Rightarrow aM^* r^{m+1} < |h^i| < a^{-1} r^{-\delta}$, where a is positive constant depending only on the constant C of (d).

Choose M^* so that

$$aM^* > M_2. \tag{3.25}$$

The relation (3.24) follows from

$$\hat{\eta}^i_{i-1}(\hat{\eta} + aM^* r^{m+2}) - \eta^i > (aM^* - M_2) r^{m+2} + O(r^{m+3}) > 0 \tag{3.26}$$

and

$$\frac{d}{dh^i}\{h^i(1 + rM_1(h))\} = 1 + O(r).$$

(ii) It is easy to see that $\tilde{T}(\hat{\eta} + rh)$ is continuously differentiable on B with derivatives

$$\tilde{T}^i_j = |\tilde{T}^i|^{-1}\left(L_k(\hat{\eta}_{i-1})\frac{\partial \hat{\eta}^k_{i-1}}{\partial h^j} - L_k(\hat{\eta}_i)\frac{\partial \hat{\eta}^k_i}{\partial h^j}\right).$$

Note that,

$$\frac{\partial \hat{\eta}^a_{i-1}}{\partial \eta^b} = \begin{cases} 0, & a, b \ge i, \\ \delta_{ab}, & a \le i - 1, \end{cases}$$

and $L_k(\hat{\eta}_{i-1}) = 0$, $k \ge i$. So $i < j \Rightarrow \tilde{T}^i_j = 0$ while

$$\tilde{T}^i_i = -r^{-1}|\tilde{T}^i|^{-1} L_i(\hat{\eta}_i). \tag{3.27}$$

Now write

$$\begin{aligned} L_i(\hat{\eta}_i) &= L_{ib}(\hat{\eta}_{i-1})(\hat{\eta}^b_i - \hat{\eta}^b_{i-1}) \\ &\quad + \sum_{k=1}^{m+1} \frac{L_{ib_1 \cdots b_k}(\hat{\eta}_{i-1})}{k!} \prod_{j=1}^{k} (\hat{\eta}^j_i - \hat{\eta}^j_{i-1}) \\ &\quad + O(|\hat{\eta}^i_{i-1} - \hat{\eta}^i|^{m+2}) \\ &= \sum_{k=1}^{m+1} P_k(rh) d^k + O(d^{m+2}) \end{aligned} \tag{3.28}$$

by (3.13), where $d = \hat{\eta}^i_{i-1} - \eta^i$ and P_k are polynomials in rh such that

$P_1(0) = 1$. Now apply (3.21) and (3.28) to (3.27) and then substitute (3.22) for d and (ii) follows.

(iii) Follows from Lemma A1 of the Appendix. □

PROOF OF THEOREM 1. By Lemma 3 formula (3.8) is valid for $(\mathbf{x}, t) \in \tilde{S}$. Moreover, from Lemma 2,

$$h^i(t) = t^i + r^{-1}P^i(rt) + O(r^{m+1}), \tag{3.29}$$

where P^i is a polynomial of degree $m+1$ in rt with no constant or linear term and coefficients which are polynomials in $L_{b_1 \cdots b_k}$, $k \le m+2$. From (3.23) and (3.29)

$$\begin{aligned} \det\left\|h_j^i(t)\right\| &= \det\left\|\tilde{T}_j^i(\hat{\eta} + rh(t))\right\|^{-1} = \prod_{i=1}^{p} \tilde{T}_i^i(\hat{\eta} + rh(t))^{-1} \\ &= \prod_{i=1}^{p}\left(1 + P^i(rh(t))\right)^{-1} + O(r^{m+1}) \\ &= 1 + V(rt) + O(r^{m+1}), \end{aligned} \tag{3.30}$$

where V is a polynomial of degree $m+1$ in rt with no constant term and coefficients which are polynomials in $L_{b_1 \cdots b_k}$, $k \le m+2$.

Moreover, from (3.29) and B_m(i),

$$\begin{aligned} \pi(\hat{\theta} + r\hat{D}^{-1}h(t)) = \pi(\hat{\theta})\Bigg(1 &+ \frac{\pi_b(\hat{\theta})}{\pi(\hat{\theta})}U^b(rt) + \cdots \\ &+ \frac{\pi_{b_1 \cdots b_{m+2}}}{\pi(\hat{\theta})}(\hat{\theta})U^{b_1}(rt) \cdots U^{b_{m+2}}(rt)\Bigg) \\ &+ O(r^{m+1}\pi(\theta)), \end{aligned} \tag{3.31}$$

where the U^b are polynomials of degree $\le m+1$ with no constant term. Substituting back (3.30) and (3.31) in (3.8) provides an approximation to the numerator in (3.8) and integrating this we get an approximation to the denominator in (3.8). Together these approximations ensure that

$$E_P\int\left|\pi_T(t|\mathbf{X}) - \phi(t)(1 + Q_m^*(rt, x, \pi))1[(t, \mathbf{X}) \in \tilde{S}]\right|dt = O(r^{m+1})$$

for a suitable Q_m^*. We get Q_m by dropping all terms of degree $m+1$ in Q_m^*. The coefficients are evidently polynomials in $L_{b_1 \cdots b_k}(\hat{\eta})$ and $\pi_{b_1 \cdots b_k}/\pi(\hat{\theta})$, $1 \le k \le m+1$. But the former are polynomials in the elements of $\hat{D}^{-1}$ which are rational functions of $L_{ij}(\hat{\theta})$. Now,

$$\begin{aligned} E_P\int\phi(t)\left[Q_m(rt, \mathbf{x}, \pi) - Q_m^*(rt, x, \pi)\right]1[(t, \mathbf{X}) \in \tilde{S}]\,dt \\ = O(r^{m+1}) \end{aligned} \tag{3.32}$$

since for $\mathbf{x} \in S$ all coefficients in both functions are bounded. Further,

$$E_P\int\pi_T(t|\mathbf{X})1((t, \mathbf{X}) \notin \tilde{S})\,dt = P[(T, \mathbf{X}) \notin \tilde{S}] = O(r^{m+1}) \tag{3.33}$$

by B_m. Finally,

$$E_P\int \phi(t)Q_m(rt, x, \pi)1(\mathbf{X} \in S, |t| \le M^* r^{m+1} \text{ or } |t| \ge r^{-\delta})\, dt = O(r^{m+1})$$

and the theorem follows. □

PROOF OF THEOREM 2 AND COROLLARY 1. Evidently since D and $\tilde{D}$ are simple transforms of T, we need merely check that the approximation to the density of D ($\tilde{D}$, respectively) obtained by applying the usual transformation formula to $\pi_m(\cdot, \mathbf{X})$ agrees with $\prod_{k=1}^{p} c_1(u^k)$ with error $O(r^{m+1})$ for $m = 1, 3$, respectively. This follows readily from Lemmas A2 and A3 in the Appendix if we identify π_m with $g(t)$ for $m = 2, 3$ and note that $R_{jj} = O(n^{-1})$. Relation (2.6) follows from Lemmas A2 and A3. Corollary 1(a) follows immediately from (2.5), while 1(b) follows from (2.6) and Lemma A4. □

PROOF OF THEOREM 3. Evidently $F_m \Rightarrow B_m$ for π satisfying (vii). It is shown in Ghosh, Sinha and Joshi (1982) and Bickel, Götze and van Zwet (1985) that the set of all such π is dense in the set of all priors under weak convergence. Now (2.9) implies that for any π concentrating on a compact, the characteristic function of T satisfies the approximation

$$\begin{aligned}
\int e^{i\nu_j t^j} p_T(t)\, dt &= \int\int e^{i\nu_j t^j} p_T(t|\theta)\pi(\theta)\, d\theta\, dt \\
&= e^{i\nu_j t^j}\phi(t)\left(1 + \sum_{k=1}^{m} r^k \int R_k(t,\theta)\pi(\theta)\, d\theta\right) dt + O(r^{m+1}) \\
&= \exp\left\{-\tfrac{1}{2}\sum_{j=1}^{p} (\nu^j)^2\right\}\left[1 + \sum_{k=1}^{m} r^k \int P_k(\nu,\theta)\pi(\theta)\, d\theta\right] \\
&\quad + O(r^{m+1}),
\end{aligned} \tag{3.34}$$

where $\exp\{-\frac{1}{2}\sum_{j=1}^{p}(\nu^j)^2\}P_k(\nu,\theta)$ is the Fourier transform of $\phi(t)R_k(t,\theta)$, so that the P_k's are also polynomials in ν. On the other hand, Theorem 1 yields

$$\begin{aligned}
&\int \exp\left\{\sum_{j=1}^{p} (\nu^j)^2\right\} p_T(t)\, dt \\
&= E_P\left[\int \exp\left\{\sum_{j=1}^{p} (\nu^j)^2\right\}\pi_m(t,\mathbf{X})1(\mathbf{X} \in S)\, dt\right] + O(r^{m+1}) \\
&= \exp\left\{-\tfrac{1}{2}\sum_{j=1}^{p} (\nu^j)^2\right\} \\
&\quad \times\left(1 + \sum_{k=1}^{m} r^k t^{b_1} \cdots t^{b_l} EQ_{m b_1 \cdots b_k}(\mathbf{X},\pi)1(\mathbf{X} \in S)\right) + O(r^{m+1}).
\end{aligned} \tag{3.35}$$

Therefore, multiplying by $\exp\{\frac{1}{2}\Sigma_{j=1}^{p}(\nu^j)^2\}$ we get

$$(3.36)\qquad \begin{aligned} 1 &+ \sum_{k=1}^{m} r^k \int P_k(\nu,\theta)\pi(\theta)\,d\theta \\ &= 1 + \sum_{k=1}^{m} r^k c_{b_1\cdots b_k}(\pi)\nu^{b_1}\cdots\nu^{b_k} + O(r^{m+1}), \end{aligned}$$

where O is now uniform for $|\nu| \le M$ by the hypothesis of Theorem 3.

Define, as usual,

$$\Delta_{b_1\cdots b_p} f(t^1,\dots,t^p) = \left(\Delta_1^{b_1}\cdots\Delta_p^{b_p}\right) f(t^1,\dots,t^p),$$

where the $b_j = 0,\dots,p$, $\Sigma_{j=1}^{p} b_j = l$ and

$$\Delta_k f = f(t^1,\dots,t^{k-1},t^k+\varepsilon,t^{k+1},\dots,t^p) - f(t^1,\dots,t^p)$$

and Δ_k^p represents an operator product. Apply $\Delta_{b_1\cdots b_p}$ to both sides of (3.36) considered as functions of ν. If $l > m$ we obtain

$$(3.37)\qquad \sum_{j=1}^{m} r^j\varepsilon^{-l}\int \Delta_{b_1\cdots b_p} P_j(\varepsilon,\theta)\pi(\theta)\,d\theta = O(r^{m+1}\varepsilon^{-l}).$$

Let $\varepsilon \downarrow 0$ more slowly than $r^{1/l}$. Then (3.37) yields

$$\int \frac{\partial^p P_k}{\partial^{b_1}u_1\cdots\partial^{b_p}u_p}(\nu,\theta)\pi(\theta)\,d\theta = 0 \quad \text{for all } \nu, \text{ for all } k \le m.$$

But by assumption the integrand is continuous in θ. Since π ranges over a dense set we conclude that the integrand vanishes identically in θ. So P_k is a polynomial of degree less than or equal to k and hence so is R_k. □

Theorem 4 and Corollary 3 follow from Theorem 3 in the same fashion as Theorem 2 and Corollary 1 follow from Theorem 1.

Acknowledgments. We thank Ole Barndorff-Nielsen, Ib Skovgaard and Peter McCullagh for some crucial references.

APPENDIX

LEMMA A1. *Suppose $f\colon C^0 \to R^p$ where C^0 is an open convex set in R^p. Suppose f is differentiable with Hessian $\dot{f}$ and*

$$(\text{A1})\qquad |\dot{f} - J| < 1,$$

where J is the identity and $|M|$ is the operator norm on matrices. Then $\dot{f}$ is nonsingular and f is 1–1.

PROOF. By (A1), $\dot{f}$ is nonsingular:

$$\dot{f}^{-1} = J - (\dot{f} - J) + (\dot{f} - J)^2 \cdots .$$

REFERENCES

BARNDORFF-NIELSEN, O. E. (1983). On a formula for the conditional distribution of the maximum likelihood estimator. *Biometrika* **70** 343–365.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *J. Roy. Statist. Soc. Ser. B* **46** 483–495.

BARNDORFF-NIELSEN, O. E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73** 307–322.

BARNDORFF-NIELSEN, O. E. and HALL, P. (1988). On the level error after Bartlett adjustment of the likelihood ratio statistic. *Biometrika* **75** 374–378.

BARTLETT, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. Roy. Soc. Lond. Ser. A* **160** 268–282.

BHATTACHARYA, R. N. and GHOSH, J. K. (1978). Validity of formal Edgeworth expansion. *Ann. Statist.* **6** 434–451.

BICKEL, P. J. (1974). Edgeworth expansions in nonparametric statistics. *Ann. Statist.* **2** 1–20.

BICKEL, P. J., GÖTZE, F. and VAN ZWET, W. R. (1985). A simple analysis of third order efficiency of estimates. In *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer* (L. Le Cam and R. A. Olshen, eds.) **2** 749–768. Wadsworth, Belmont, Calif.

BILLINGSLEY, P. (1961). *Statistical Inference for Markov Processes*. Univ. Chicago Press, Chicago.

BOX, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika* **36** 317–346.

CHANDRA, T. and GHOSH, J. K. (1979). Valid asymptotic expansion for the likelihood ratio statistic and other perturbed χ^2 variables. *Sankhyā Ser. A* **41** 22–47.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.

EFRON, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika* **72** 45–58.

GHOSH, J. K., SINHA, B. K. and JOSHI, S. M. (1982). Expansions for posterior probability and integrated Bayes risk. In *Statistical Decision Theory and Related Topics III* (S. S. Gupta and J. O. Berger, eds.) **1** 403–456. Academic, New York.

GÖTZE, F. and HIPP, C. (1978). Asymptotic expansions under moment conditions. *Z. Wahrsch Verw. Gebiete* **42** 67–87.

LAWLEY, D. N. (1956). A general method for approximating to the distribution of the likelihood ratio criteria. *Biometrika* **43** 295–303.

PFANZAGL, J. (1974). Asymptotically optimum estimation and test procedures. In *Proc. Prague Symp. on Asymptotic Statistics* (J. Hájek, ed.) **1** 201–272. Charles Univ., Prague.

STEIN, C. (1985). On the coverage probability of confidence sets based on a prior distribution. *Sequential Meth. Statist.: Banach Center Publication* **16** 485–514.

WALD, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* **54** 426–482.

WELCH, B. N. and PEERS, B. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. Ser. B* **35** 318–329.

WILKS, S. S. (1938). The large sample distribution of the likelihood ratio statistic for testing composite hypotheses. *Ann. Math. Statist.* **9** 60–62.

DEPARTMENT OF STATISTICS
STATISTICAL LABORATORY
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720

INDIAN STATISTICAL INSTITUTE
203 BARRACKPORE TRUNK RD.
CALCUTTA 700 0-35
INDIA

J.K. Ghosh, S.K. Mitra, K.R. Parthasarathy and B.L.S. Prakasa Rao (*eds.*),
Statistics and Probability : A Raghu Raj Bahadur Festschrift
Published by Wiley Eastern Limited, Publishers, (1993), 83 – 96

Asymptotic Distribution of the Likelihood Ratio Statistic in a Prototypical Non Regular Problem

P. BICKEL[1]
University of California at Berkeley, Berkeley

and

H. CHERNOFF[2]
Harvard University, Cambridge

Abstract

This paper addresses the asymptotic behavior of $M_n^* = \sup_t S_n^*(t)$ where

$$\begin{aligned} S_n^*(t) &= n^{-1/2}\sum_{i=1}^{n} y^*(X_i,t) \\ y^*(x,t) &= (e^{tx-t^2/2} - 1 - tx)/(e^{t^2} - 1 - t^2)^{1/2} \end{aligned}$$

and $X_1, X_2, \ldots, X_n$ are i.i.d. $N(0,1)$ random variables.

[1]Supported in part by NSF Grant DMS-8505550 at the Mathematical Sciences Research Institute

[2]Supported in part by NSF Grant DMS-8505550 and ONR contract N00014-91-J1005.

AMS Subject Classifications: Primary 60G70; secondary 60G05, 62M99.

Key words and Phrases: Asymptotic distribution, supremum of stochastic process, Slepian theorem, Iterated logarithm, Kolmogorov bound, Hungarian construction, likelihood-ratio tests, mixtures of distributions.

It will be shown that

$$M_n^* = (\log_2 n)^{1/2} + (V_n - \log(\sqrt{2\pi}))(\log_2 n)^{-1/2}$$

where $\log_2 n = \log(\log n)$ and, as $n \to \infty$,

$$P\{V_n \leq v\} \to \exp(-e^{-v}).$$

This result gives the asymptotic behavior of the likelihood-ratio test statistic for testing whether a mixture of two normal distributions with specified variance is simply a pure normal with that variance. We expect that the analysis for this problem will be relevant to a class of testing problems, including mixture problems and change point problems, characterized by a singularity in the natural parametrization when the null hypothesis is true, that leads to a loss of indentification under the null hypothesis.

1 Introduction

This paper addresses the asymptotic behavior of $M_n^* = \sup_t S_n^*(t)$ where

$$S_n^*(t) = n^{-1/2} \sum_{i=1}^{n} y^*(X_i, t) \tag{1}$$

$$y^*(x,t) = (e^{tx-t^2/2} - 1 - tx)/(e^{t^2} - 1 - t^2)^{1/2} \tag{2}$$

and $X_1, X_2, \ldots, X_n$ are i. i. d. $N(0,1)$ random variables. It will be shown that

$$M_n^* = (\log_2 n)^{1/2} + (V_n - \log(\sqrt{2\pi}))(\log_2 n)^{-1/2} \tag{3}$$

where $\log_2 n = \log(\log n)$ and, as $n \to \infty$,

$$P\{V_n \leq v\} \to \exp(-e^{-v}). \tag{4}$$

Hartigan (1984) has shown that the logarithm of the likelihood-ratio test statistic for the test that a mixture of two normal distributions with known variance is a pure normal with that variance, can be approximated by an expression stochastically equal to M_n^{*2} and that $M_n^* \to \infty$ in probability. His proof uses the fact that the $y^*(X_i, t)$ have mean 0 and variance 1, and hence $S_n^*(t)$ is asymptotically normal with mean 0 and variance 1 for each value of t. The covariances of S_n^* at s and t become small for s and t far enough apart, and hence M_n^* relates to the maximum of a large number of almost independent normals, and approaches infinity. He conjectures correctly that $M_n^* = O((\log_2 n)^{1/2})$.

There is a large class of problems, including mixture problems and change point problems, that are characterized by a natural parameterization which degenerates under the null hypothesis (See Ghosh and Sen (1985)). For example, in Hartigan's problem, the underlying parameters are α, μ_1 and μ_2, the mixture proportion and the two means. A single distribution under the null hypothesis, e.g. $N(0,1)$, can be represented by infinitely many combinations of these parameters. In other words, there is a loss of identification under the null hypothesis.

Heuristically, what happens in such problems is the following (compare Chernoff (1954).) We suppose we have $X_1, \dots, X_n$ i.i.d. $p(\cdot, \theta)$, $\theta = (\phi, \psi)$, Euclidean. Suppose also that ϕ is unidentifiable if $\psi = 0$, $p(\cdot, \phi, 0) = p(\cdot, 0, 0)$. We expect that, for $\psi = O(n^{-1/2})$,

$$\sum_{j=1}^{n} \log \frac{p(X_i, \theta)}{p(X_i, \phi, 0)} \simeq \Psi^T \sum_{i=1}^{n} T(X_i, \phi)$$
$$-\frac{n}{2} \Psi^T \sum(\phi) \Psi + o_p(1)$$

where $E_0 T = 0$, $\mathrm{Var}_0 T = \sum(\phi)$, and hence that,

$$2 \sup_{\phi, \psi \neq 0} \sum_{i=1}^{n} \log \frac{p(X_i, \theta)}{p(X_i, \phi, 0)} = \sup_{\phi} \|Z_n(\phi)\|^2 + o_p(1)$$

where

$$Z_n(\phi) = n^{-1/2} \sum_{i=1}^{n} \sum\nolimits^{-1/2}(\phi) T(X_i, \phi).$$

It is reasonable to expect that $\sup_\phi \|Z_n(\phi)\|^2$ behaves to first order like $\sup_\phi \|Z(\phi)\|^2$ where Z is Gaussian, $EZ(\phi) = 0$ and $\mathrm{Var} Z(\phi) = I$ (the identity). One possibility now is that $P[\sup_\phi Z(\phi) < \infty] = 1$. We expect this to happen if the domain of ϕ is compact, for instance, if X is distributed as $(1-\psi)\ \mathrm{Bin}(m, \frac{1}{2}) + \psi\ \mathrm{Bin}(m, \phi)$, a case considered in Chernoff and Lander (1989). It can even happen if ϕ is unbounded. Take X distributed as bivariate $\mathcal{N}(\psi, \phi\psi, I)$, in which case, $\sup_\phi \|Z_n(\phi)\|^2$ is χ_2^2, rather than the naively expected χ_1^2. However, in the normal mixture problem, $\sup_\phi \|Z_n(\phi)\|^2 = \infty$. We may still hope, however, that $P[\sup_\phi \|Z_n(\phi)\|^2 < \infty] = 1$. Suppose further, $\sup_\phi \|Z_n(\phi)\|^2 = \sup\{\|Z_n(\phi)\|^2 : \phi \in K_n\}$ plus lower order terms where K_n is compact, K_n tends, as $n \to \infty$, to the range of ϕ (which can be a cone rather than the whole Euclidean space). Then we may expect that the theory of extrema of Gaussian processes — see Leadbetter, Lindgren and Rootzen (1983) will be a guide to higher order behavior of $\sup_\phi \|Z_n(\phi)\|^2$. That is the prescription that we carry out in this paper for the special case of normal mixtures. We expect the approach to be applicable to most other mixture and change point problems.

2 Outline

Instead of attacking $S_n^*(t)$ directly, we will first deal with the simpler expression

$$S_n(t) = n^{-1/2} \sum_{i=1}^{n} (e^{tX_i - t^2/2} - 1) e^{-t^2/2} \tag{5}$$

with supremum

$$M_n = sup_t S_n(t). \tag{6}$$

Then

$$\begin{aligned} S_n(t) &= n^{1/2} \int_0^1 y(x,t)[dF_n(u) - dF(u)] \\ &= \int_0^1 y(x,t)\, dB_n(u) \end{aligned} \tag{7}$$

where

$$y(x,t) = e^{tx-t^2} \tag{8}$$

and

$$\begin{gathered} u = \Phi(x) = \textstyle\int_{-\infty}^{x} \phi(v)\, dv = 1 - \Phi^c(x), \\ \phi(v) = (2\pi)^{-1/2} \exp(-v^2/2), \\ F_n(u) = \text{sample c. d. f. of } U_i = \Phi(X_i) = 1 - U_i^c, \\ F(u) = u = 1 - u^c, \end{gathered}$$

and

$$B_n(u) = n^{1/2}[F_n(u) - F(u)]. \tag{9}$$

We will relate $S_n(t)$ to the Gaussian Process

$$S_0(t) = \int_0^1 y(x,t) dB_0(u) \tag{10}$$

where B_0 is the Brownian Bridge. The Hungarian Construction (Komlos, Major, Tusnady, (1975)) gives us

$$sup_{0 \leq u \leq 1} |B_n(u) - B_0(u)| = O_p(n^{-1/2} \log n) \tag{11}$$

on a suitable probability space.

Both S_0 and S_n are zero mean stochastic processes with common covariance function

$$\rho(s,t) = \exp(-(s-t)^2/2) - \exp(-s^2/2 - t^2/2).$$

By adjoining to S_0 and S_n, the relatively trivial additional term $\tilde{X}e^{-t^2/2}$ where $\tilde{X}$ is independent of S_0 and S_n and $\mathcal{L}(\tilde{X}) = N(0,1)$, we have the processes

$$\tilde{S}_0(t) = S_0(t) + \tilde{X}e^{-t^2/2} \tag{12}$$

and

$$\tilde{S}_n(t) = S_n(t) + \tilde{X}e^{-t^2/2} \tag{13}$$

with common covariance function

$$\tilde{\rho}(s,t) = e^{-(s-t)^2/2} \tag{14}$$

Since $\tilde{S}_0$ is a Gaussian and stationary process its maximal behavior is well known (see Leadbetter, 1983) i.e.,

$$\begin{aligned}\tilde{M}(T) &= sup_{-T\le t\le T}\tilde{S}_0(t)\\ &= (2\log 2T)^{1/2} + (V_T - \log(2\pi))(2\log 2T)^{-1/2}\end{aligned} \tag{15}$$

where, as $T \to \infty$,

$$P\{V_T \le v\} \to \exp(-e^{-v}). \tag{16}$$

Furthermore, the location of the value of t for which $\tilde{S}_0$ attains its maximum is uniformly distributed over the range $(-T, T)$.

The main idea of the derivation is to show that $S_n(t)$ behaves like $S_0(t)$ and $\tilde{S}_0(t)$ for $|t| \le \sqrt{\log n/2}$, and becomes relatively small for $|t| \ge \sqrt{\log n/2}$.

For $t > 0$, we shall decompose each of the integrals S_0 and S_n into two parts in one of two ways. Then, for $i = 1, 2$

$$\begin{aligned}S^u_{ni}(t) &= \int_{x>x_{ni}} y(x,t)dB_n(u),\\ S^u_{0i}(t) &= \int_{x>x_{ni}} y(x,t)dB_0(u),\\ S^l_{ni}(t) &= \int_{x\le x_{ni}} y(x,t)dB_n(u),\\ S^l_{0i}(t) &= \int_{x\le x_{ni}} y(x,t)dB_0(u)\end{aligned}$$

We will also find it convenient to define, for $i = 1, 2$,

$$\begin{aligned}\tilde{S}^u_{0i}(t) &= \tilde{X}e^{-t^2/2}\Phi^c(x_{ni}-t) + S^u_{0i}(t)\\ \tilde{S}^u_{ni}(t) &= \tilde{X}e^{-t^2/2}\Phi^c(x_{ni}-t) + S^u_{ni}(t)\\ \tilde{S}^l_{0i}(t) &= \tilde{X}e^{-t^2/2}\Phi(x_{ni}-t) + S^l_{0i}(t)\\ \tilde{S}^l_{ni}(t) &= \tilde{X}e^{-t^2/2}\Phi(x_{ni}-t) + S^l_{ni}(t)\end{aligned}$$

since the corresponding covariance functions are

$$\tilde{\rho}^u_i(s,t) = \Phi^c(x_{ni} - (s+t))e^{-(s-t)^2/2} \tag{17}$$

for $\tilde{S}^u_{0i}$ and $\tilde{S}^u_{ni}$, and

$$\tilde{\rho}^l_i(s,t) = \Phi(x_{ni} - (s+t))e^{-(s-t)^2/2} \tag{18}$$

for $\tilde{S}^l_{0i}$ and $\tilde{S}^l_{ni}$.

Our main immediate goal is to show that M_n is stochastically asymptotically equivalent to $\tilde{M}(\sqrt{\log n/2})$. The argument involves several ranges of t and two different levels of x_{ni}, and five basic tools or established theorems. The levels of x_{ni} and the crucial t values are given by

$$\begin{aligned}
x_{n1}^2 &= 2\log n - 4\log_2 n \\
x_{n2}^2 &= 2\log n - 2\log_2 n \\
t_{n0} &= (2\log_3 n)^{1/2} \\
t_{n1} &= x_{n1}/2 - 2(\log_2 n)^{1/2} \\
t_{n2} &= x_{n1}/2 - 2(\log_3 n)^{1/2} \\
t_{n3} &= x_{n2}/2 + 2(\log_3 n)^{1/2} \\
t_{n4} &= x_{n2}/2 + 2(\log_2 n)^{1/2} \\
t_{n5} &= (\log n)^{1/2},
\end{aligned}$$

where $\log_3 n \equiv \log(\log_2 n)$. Note that $0 < x_{n2} - x_{n1} = o(1)$.

We shall abbreviate the basic tools with letters. Thus the Hungarian Construction, referred to previously, will be labeled H. The law of the iterated logarithm will be labeled I. Slepian's theorem states that if $\rho_1(s,t)$ and $\rho_2(s,t)$ are the autocovariances for two Gaussian processes with mean 0 and variance 1, and $\rho_1(s,t) \geq \rho_2(s,t)$ for all s,t, then the supremum of the first process is stochastically smaller than the supremum of the second, (see Leadbetter, 1983), and will be labeled S.

Another basic tool, to be labeled T for Tail, is the fact that

$$\sup_{0\leq u\leq 1} [F_n(u)/F(u)] = O_p(1) \tag{19}$$

and

$$\sup_{0\leq u\leq 1} [(1-F_n(u))/(1-u)] = O_p(1) \tag{20}$$

Finally, a fifth tool to be labeled K is the Kolmogorov Bound, (see Billingsley (1968)), which states that if $Z(t)$ is a stochastic process which satisfies

$$E[Z(s) - Z(t)]^2 \leq c(s-t)^2, \tag{21}$$

for $0 \leq s \leq t \leq 1$, then

$$P[\sup_{0\leq t\leq 1} |Z(t) - Z(0)| \geq z] \leq Kc/z^2 \tag{22}$$

where K is an absolute constant.

The main argument will show that $S_n(t) - S_0(t) = o_p(\log_2 n)^{-1/2}$ for $|t| \leq t_{n1}$. Since the difference between $\sup_{|t|\leq t_{n1}} S_0(t)$ and $\sup_{|t|\leq t_{n1}} \tilde{S}_0(t)$ is negligible, that will establish that $\sup_{|t|\leq t_{n1}} S_n(t)$ behaves asymptotically like $\tilde{M}(\sqrt{\log n/2})$. But we want this result for $\sup_t S_n(t)$. To achieve this it suffices then to show that $\sup_{|t|>t_{n1}} S_n(t) = o_p(\log_2 n)^{1/2}$, and hence the supremum over all t is achieved for $|t| \leq t_{n1}$ with probability approaching one. Finally the difference between M_n and M_n^* will be shown to be negligible.

We now outline in more detail the arguments used to show that $S_n(t) - S_0(t) = o_p(\log_2 n)^{-1/2}$ for $0 \leq t \leq t_{n1}$ and $S_n(t) = o_p(\log_2 n)^{1/2}$ for $t \geq t_{n1}$. Recall that $S_n(t) = S_{n1}^l(t) + S_{n1}^u(t) = S_{n2}^l(t) + S_{n2}^u(t)$ and that $S_0(t)$ can be decomposed similarly.

(1) $0 \leq t \leq t_{n1}$

$H1$: $S_{n1}^l(t) - S_{01}^l(t) = o_p(\log_2 n)^{-1/2}$

$I1$: $S_{01}^u(t) = o_p(\log_2 n)^{-1/2}$

$T1$: $S_{n1}^u(t) = o_p(\log_2 n)^{-1/2}$

(2) $t_{n1} \leq t \leq t_{n2}$

$H2$: $S_{n1}^l(t) - S_{01}^l(t) = o_p(\log_2 n)^{1/2}$

$S2$: $S_{01}^l(t) = o_p(\log_2 n)^{1/2}$

$K2$: $S_{n1}^u(t) = o_p(\log_2 n)^{1/2}$

(3) $t_{n2} \leq t \leq t_{n3}$

$K3$: $S_n(t) = o_p(\log_2 n)^{1/2}$

(4) $t_{n3} \leq t \leq t_{n4}$

$K4$: $S_{n2}^l(t) = o_p(\log_2 n)^{1/2}$

$T4$: $S_{n2}^u(t) = o_p(\log_2 n)^{1/2}$

(5) $t_{n4} \leq t \leq t_{n5}$

$K5$: $S_{n2}^l(t) = o_p(\log_2 n)^{1/2}$

$T5$: $S_{n2}^u(t) = o_p(\log_2 n)^{1/2}$

(6) $t_{n5} \leq t$

$T6$: $S_n(t) = o_p(\log_2 n)^{1/2}$

3 The Kolmogorov Bound argument

First we note a simple corollary of the Kolmogorov Bound result. If the interval over which Z is defined, is replaced by one of length L, then the bound cK/z^2, for the probability of the maximum deviation, is replaced by $cK(L/z)^2$. Consequently, we have $\sup(|Z(s) - Z(t)|) = (Lc^{1/2})O_p(1)$.

3.1 $K3$ Since $\tilde{\rho}(s,t) = \exp[-(s-t)^2/2]$,

$$\begin{aligned} E\{[\tilde{S}_n(s) - \tilde{S}_n(t)]^2\} &= 2(1 - e^{-(s-t)^2/2}) \\ &\leq (s-t)^2 \end{aligned}$$

and it follows that

$$\begin{aligned} \sup_{t_{n2} \leq t \leq t_{n3}} |\tilde{S}_n(t) - \tilde{S}_n(t_{n2})| &= O_p(t_{n3} - t_{n2}) \\ &= o_p(\log_2 n)^{1/2}. \end{aligned}$$

Then

$$\begin{aligned}\sup_{t_{n2}\le t\le t_{n3}} S_n(t) &\le S_n(t_{n2}) + \sup_{t_{n2}\le t\le t_{n3}} |\tilde{S}_n(t) - \tilde{S}_n(t_{n2})| \\ &\quad + \sup_{t_{n2}\le t\le t_{n3}} |e^{-t^2/2} - e^{-t_{n2}^2/2}| O_p(1) \\ &= O_p(1) + o_p(\log_2 n)^{1/2} + o_p(1) \\ &= o_p(\log_2 n)^{1/2} \end{aligned} \tag{23}$$

3.2 $K2$ From Equation 17 it follows that

$$\begin{aligned} E\{[\tilde{S}^u_{n1}(s) - \tilde{S}^u_{n1}(t)]^2\} &= \Phi^c(x_{n1} - 2s) + \Phi^c(x_{n1} - 2t) \\ &\quad -2\Phi^c(x_{n1} - (t+s)) + 2\Phi^c(x_{n1} - (t+s))[1 - e^{-(s-t)^2/2}] \end{aligned}$$

For $0 \le t \le s \le x_{n1}/2 - 1$,

$$\begin{aligned} \Phi^c(x_{n1} - 2s) + \Phi^c(x_{n1} - 2t) - 2\Phi^c(x_{n1} - (t+s)) &\le \\ (x_{n1} - 2s)\phi(x_{n1} - 2s)(s-t)^2 \end{aligned}$$

and thus

$$E\{[\tilde{S}^u_{n1}(s) - \tilde{S}^u_{n1}(t)]^2\} \le \{\Phi^c(x_{n1} - 2s) + (x_{n1} - 2s)\phi(x_{n1} - 2s)\}(s-t)^2,$$

and for $t_{n1} \le t \le s \le t_{n2}$, the coefficient of $(s-t)^2$ is bounded by $O[(x_{n1} - 2t_{n2})\phi(x_{n1} - 2t_{n2})]$. It follows that

$$\begin{aligned}\sup_{t_{n1}\le t\le t_{n2}} S^u_{n1}(t) &\le O_p(1) + O_p(t_{n2} - t_{n1}) O[(x_{n1} - 2t_{n2})\phi(x_{n1} - 2t_{n2})]^{1/2} \\ &= O_p(1) + O_p((\log_2 n)^{1/2}) o(1) \\ &= o_p((\log_2 n)^{1/2}). \end{aligned} \tag{24}$$

3.3 $K4$ Essentially the same argument as that used for K2 applies to $S^l_{n2}(t)$. Here $\Phi^c(x_{n1} - 2t)$ in the proof for K2 is replaced by $\Phi(x_{n2} - 2t)$ and the argument $x_{n2} - 2t$ is negative.

3.4 $K5$ Here again, the same argument applies, except that in place of Equation 24, we have

$$\begin{aligned}\sup_{t_{n4}\le t\le t_{n5}} S^{\ell}_{n2}(t) &= O_p(1) + O_p(\log n)^{1/2} O[(x_{n2} - 2t_{n4})\phi(x_{n2} - 2t_{n4})]^{1/2} \\ &= O_p(1) + O_p(\log n)^{1/2} O[(\log_2 n)^{1/2}\ e^{-16\log_2 n/2}]^{1/2} \\ &= O_p(1) \end{aligned} \tag{25}$$

4 The Slepian argument, $S2$

The process $\tilde{S}^l_{01}(t)[\Phi((x_{n1} - 2t))]^{-1/2}$ has covariance function

$$\frac{\tilde{\rho}^l_{01}(s,t)}{[\tilde{\rho}^l_{01}(s,s)\tilde{\rho}^l_{01}(t,t)]^{1/2}} = \frac{\Phi(x_{n1} - (t+s))}{[\Phi(x_{n1} - 2s)\Phi(x_{n1} - 2t)]^{1/2}} e^{-(s-t)^2/2}.$$

It is easy to see that $\log \dot\Phi(x)$ is concave, and hence this covariance function is no less that $e^{-(s-t)^2/2}$, and, by Slepian's theorem

$$\begin{aligned}\sup_{t_{n1}\le t\le t_{n2}} \tilde S^l_{01}(t) &= O_p(\log(t_{n2}-t_{n1}))^{1/2}[\Phi(x_{n1}-2t_{n1})]^{1/2}\\ &= O_p[(\log_2 n)^{1/4}]O(1),\end{aligned}$$

and

$$\sup_{t_{n1}\le t\le t_{n2}} S^l_{01}(t) = o_p(\log_2 n)^{1/2}. \tag{26}$$

5 The Hungarian Construction argument, $H1, H2$

The following argument applies for $0 \le t \le t_{n2}$, i. e., for both $H1$ and $H2$. We have

$$S^l_{n1}(t) - S^l_{01}(t) = \int_{x\le x_{n1}} e^{tx-t^2}[dB_n(u) - dB_0(u)].$$

and after integration by parts,

$$\begin{aligned}|S^l_{n1}(t) - S^l_{01}(t)| &\le \sup_u |B_n(u) - B_0(u)|e^{tx_{n1}-t^2}\\ &= O_p(n^{-1/2}\log n)\exp[-(t-x_{n1}/2)^2 + x^2_{n1}/4]\\ &= O_p(n^{-1/2}\log n)(\log_2 n)^{-4}(n^{1/2}(\log n)^{-1})\\ &= o_p(\log_2 n)^{-1/2}. \end{aligned}\tag{27}$$

6 Iterated Logarithm argument, $I1$

The Law of the Iterated Logarithm implies that

$$B_0(u) = O_p(1)[(1-u)\log_2(1-u)^{-1}]^{1/2}. \tag{28}$$

Integrating by parts, and concentrating on $0 \le t \le t_{n1}$, we have

$$\begin{aligned}S^u_{01}(t) &= \int_{x>x_{n1}} e^{tx-t^2}\,dB_0(u)\\ &= -\int_{x>x_{n1}} B_0(u)\,d(e^{tx-t^2}) - B_0[\Phi(x_{n1})]e^{tx_{n1}-t^2}. \end{aligned}\tag{29}$$

$$\begin{aligned}B_0[\Phi(x_{n1})]e^{tx_{n1}-t^2} &= O_p(1)[\Phi^c(x_{n1})\log_2[\Phi^c(x_{n1})]^{-1}]^{1/2}e^{tx_{n1}-t^2}\\ &= O_p(1)[x_{n1}^{-1/2}\exp(-x^2_{n1}/4)(\log x_{n1})^{1/2}e^{tx_{n1}-t^2}\\ &= O_p(1)\exp(-(t-x_{n1}/2)^2)\left[\frac{\log x_{n1}}{x_{n1}}\right]^{1/2}\end{aligned}$$

$$\begin{aligned}\sup_{0\le t\le t_{n1}} |B_0[\Phi(x_{n1})]e^{tx_{n1}-t^2}| &= O_p(1)(\log n)^{-4}\left[\frac{\log_2 n}{\sqrt{\log n}}\right]^{1/2} \\ &= o_p(\log_2 n)^{-1/2}. \qquad (30)\end{aligned}$$

Also

$$\begin{aligned}\int_{x>x_{n1}} B_0(u)\,d(e^{tx-t^2}) &= t\int_{x_{n1}}^{\infty} B_0(u)e^{tx-t^2}\,dx \\ &= tO_p(1)\int_{x_{n1}}^{\infty}\frac{\exp(-x^2/4)}{x^{1/2}}(\log x)^{1/2}e^{tx-t^2}\,dx \\ &\quad \text{by (28) and the usual estimate for } \Phi^c \\ &= tO_p(1)\int_{x_{n1}}^{\infty}\left(\frac{\log x}{x}\right)^{1/2}\exp[-(t-x/2)^2]\,dx \\ &= O_p(1)t\left[\frac{\log x_{n1}}{x_{n1}}\right]^{1/2}\Phi^c\left(\frac{x_{n1}-2t}{\sqrt{2}}\right).\end{aligned}$$

$$\begin{aligned}\sup_{0\le t\le t_{n1}} \left|\int_{x>x_{n1}} B_0(u)\,d(e^{tx-t^2})\right| &= O_p(1)t_{n1}\left(\frac{\log x_{n1}}{x_{n1}}\right)^{1/2}\Phi^c\left(\frac{x_{n1}-2t_{n1}}{\sqrt{2}}\right) \\ &= O_p(1)(\log n)^{1/2}\frac{(\log_2 n)^{1/2}}{(\log n)^{1/4}}\frac{(\log n)^{-4}}{(\log_2 n)^{1/2}} \\ &= o_p(\log_2 n)^{-1/2}. \qquad (31)\end{aligned}$$

It follows that

$$\sup_{0\le t\le t_{n1}} |S_{01}^u(t)| = o_p(\log_2 n)^{-1/2} \qquad (32)$$

7 Tail argument

7.1 *T*1 We have

$$S_{n1}^u(t) = \int_{x>x_{n1}} e^{tx-t^2}n^{1/2}[dF_n(u)-dF(u)].$$

Since $F_n^c(u) = O_p(1)(1-u)$ for $0\le u\le 1$

$$\begin{aligned}S_{n1}^u(t) &= O_p(1)n^{1/2}\int_{x>x_{n1}} e^{tx-t^2}\,du \\ &= O_p(1)n^{1/2}e^{-t^2/2}\int_{x_{n1}}^{\infty} e^{-(x-t)^2/2}\,dx \\ &= O_p(1)n^{1/2}e^{-t^2/2}\Phi^c(x_{n1}-t). \qquad (33)\end{aligned}$$

Then

$$\sup_{0\le t\le t_{n1}} |S_{n1}^u(t)| = O_p(1)n^{1/2}\frac{e^{-(x_{n1}-t_{n1})^2/2}}{x_{n1}-t_{n1}}e^{-t_{n1}^2/2}$$

$$
\begin{aligned}
&= O_p(1)n^{1/2}\exp[-x_{n1}^2/4-(t_{n1}-x_{n1}/2)^2]/(x_{n1}-t_{n1})\\
&= O_p(1)n^{1/2}\frac{n^{-1/2}(\log n)}{(\log n)^{1/2}(\log n)^4}\\
&= o_p(\log_2 n)^{-1/2}. \qquad (34)
\end{aligned}
$$

7.2 $T4, T5$ The same argument yields

$$\sup_{t_{n3}\le t\le t_{n5}} |S_{n2}^u(t)| = O_p(1)n^{1/2}\sup_{t\ge t_{n3}} \Phi^c(x_{n2}-t)e^{-t^2/2}.$$

But it is not hard to show that $\Phi^c(x_{n2}-t)e^{-t^2/2}$ attains its maximum value for $t \approx x_{n2}/2+1/x_{n2}$, and that maximum value satisfies

$$\sup_t \Phi^c(x_{n2}-t)e^{-t^2/2} = \sqrt{2/\pi}e^{-x_{n2}^2/4}(x_{n2})^{-1}(1+o(1)).$$

Thus

$$
\begin{aligned}
\sup_{t\ge t_{n3}} S_{n2}^u(t) &= O_p(1)n^{1/2}\frac{n^{-1/2}(\log n)^{1/2}}{(\log n)^{1/2}}\\
&= O_p(1). \qquad (35)
\end{aligned}
$$

7.3 $T6$ Replacing x_{n1} in Equation 33 by $-\infty$, we have

$$
\begin{aligned}
\sup_{t_{n5}\le t} S_n(t) &= O_p(1)n^{1/2}\ e^{-t_{n5}^2/2}\\
&= O_p(1). \qquad (36)
\end{aligned}
$$

8 Assembly

We shall derive our result for $M(t) = \sup_t S_n(t)$ by assembling the results of Sections 3 to 7 and showing that $\sup_t S_n(t)$, $\sup_{|t|\le t_{n1}} S_n(t)$, and $\sup_{|t|\le t_{n1}} S_0(t)$ are stochastically equivalent to $\tilde{M}(t_{n1}) + o_p(\log_2 n)^{-1/2}$, where we recall that $\tilde{M}(T) = \sup_{|t|\le T}\tilde{S}_0(t)$. But first we shall show that

$$\sup_{|t|\le t_{n1}} S_0(t) = \tilde{M}(t_{n1}) + o_p(\log_2 n)^{-1/2}. \qquad (37)$$

The difficulty in this first step is that the difference $\tilde{X}e^{-t^2/2}$ between $S_0(t)$ and $\tilde{S}_0(t)$ is of the order $O_p(1)$ for t small. Thus it is useful to show that the region $|t| \ge t_{n0} = (2\log_3 n)^{1/2}$, where $\tilde{X}e^{-t^2/2} = o_p(\log_2 n)^{-1/2}$, is very likely to contain the maximizing values of t in the above suprema.

Because the location of the maximizing value of the stationary process $\tilde{S}_0(t)$ is uniformly distributed in the range of t considered, the probability that the maximizing value of t for the range $|t| \le t_{n1}$ is located within

$$
\begin{aligned}
&= O_p(1)n^{1/2}\exp[-x_{n1}^2/4 - (t_{n1} - x_{n1}/2)^2]/(x_{n1} - t_{n1}) \\
&= O_p(1)n^{1/2}\frac{n^{-1/2}(\log n)}{(\log n)^{1/2}(\log n)^4} \\
&= o_p(\log_2 n)^{-1/2}. \qquad (34)
\end{aligned}
$$

7.2 $T4, T5$ The same argument yields

$$\sup_{t_{n3}\le t\le t_{n5}} |S_{n2}^u(t)| = O_p(1)n^{1/2} \sup_{t\ge t_{n3}} \Phi^c(x_{n2} - t)e^{-t^2/2}.$$

But it is not hard to show that $\Phi^c(x_{n2}-t)e^{-t^2/2}$ attains its maximum value for $t \approx x_{n2}/2 + 1/x_{n2}$, and that maximum value satisfies

$$\sup_t \Phi^c(x_{n2} - t)e^{-t^2/2} = \sqrt{2/\pi}e^{-x_{n2}^2/4}(x_{n2})^{-1}(1 + o(1)).$$

Thus

$$
\begin{aligned}
\sup_{t\ge t_{n3}} S_{n2}^u(t) &= O_p(1)n^{1/2}\frac{n^{-1/2}(\log n)^{1/2}}{(\log n)^{1/2}} \\
&= O_p(1). \qquad (35)
\end{aligned}
$$

7.3 $T6$ Replacing x_{n1} in Equation 33 by $-\infty$, we have

$$
\begin{aligned}
\sup_{t_{n5}\le t} S_n(t) &= O_p(1)n^{1/2}\ e^{-t_{n5}^2/2} \\
&= O_p(1). \qquad (36)
\end{aligned}
$$

8 Assembly

We shall derive our result for $M(t) = \sup_t S_n(t)$ by assembling the results of Sections 3 to 7 and showing that $\sup_t S_n(t)$, $\sup_{|t|\le t_{n1}} S_n(t)$, and $\sup_{|t|\le t_{n1}} S_0(t)$ are stochastically equivalent to $\tilde{M}(t_{n1}) + o_p(\log_2 n)^{-1/2}$, where we recall that $\tilde{M}(T) = \sup_{|t|\le\ T} \tilde{S}_0(t)$. But first we shall show that

$$\sup_{|t|\le t_{n1}} S_0(t) = \tilde{M}(t_{n1}) + o_p(\log_2 n)^{-1/2}. \qquad (37)$$

The difficulty in this first step is that the difference $\tilde{X}e^{-t^2/2}$ between $S_0(t)$ and $\tilde{S}_0(t)$ is of the order $O_p(1)$ for t small. Thus it is useful to show that the region $|t| \ge t_{n0} = (2\log_3 n)^{1/2}$, where $\tilde{X}e^{-t^2/2} = o_p(\log_2 n)^{-1/2}$, is very likely to contain the maximizing values of t in the above suprema.

Because the location of the maximizing value of the stationary process $\tilde{S}_0(t)$ is uniformly distributed in the range of t considered, the probability that the maximizing value of t for the range $|t| \le t_{n1}$ is located within

$[-t_{n0}, t_{n0}]$ approaches zero. Moreover the fact that $\tilde{M}(t_{n0}) = O_p(\log_4 n)^{1/2}$ while $\tilde{M}(t_{n1}) = (\log_2 n)^{1/2} + o_p(1)$ implies that $\sup_{|t|\leq t_{n0}} S_0(t) = o_p(\log_2 n)^{1/2}$ and the supremum of $S_0(t)$ over $[-t_{n1}, t_{n1}]$ will take place for $|t| \geq t_{n0}$ where $S_0(t)$ and $\tilde{S}_0(t)$ differ by $o_p(\log_2 n)^{-1/2}$, thus establishing Equation 37.

Now we proceed to combine the results of Sections 3 to 7. It is clear that

$$\sup_{t\geq t_{n1}} S_n(t) = o_p(\log_2 n)^{1/2}. \tag{38}$$

For $0 \leq t \leq t_{n1}$,

$$\begin{aligned} S_n(t) &= S^u_{n1}(t) + [S^l_{n1}(t) - S^l_{01}(t)] - S^u_{01}(t) + S_0(t) \\ &= o_p(\log_2 n)^{-1/2} + S_0(t) \end{aligned} \tag{39}$$

By symmetry, Equation 39 holds also for $-t_{n1} \leq t \leq 0$, and Equation 38 holds for the supremum over $t \leq -t_{n1}$. Thus

$$\sup_t S_n(t) = \sup_{|t|\leq t_{n1}} S_0(t) + o_p(\log_2 n)^{-1/2}$$

is stochastically equivalent to $\tilde{M}(t_{n1}) + o_p(\log_2 n)^{-1/2}$, and therefore also to $\tilde{M}(\sqrt{\log n}/2) + o_p(\log_2 n)^{-1/2}$. It follows that

$$\sup_t S_n(t) = (\log_2 n)^{1/2} + (V - \log(\sqrt{2}\pi) + o_p(1))(\log_2 n)^{-1/2}. \tag{40}$$

This is the desired result for $S_n(t)$.

In the next section, where we extend our result to hold for S^*_n, we will use the fact that the the supremum of $S_n(t)$ is attained for $|t| > t_{n0}$.

9 Extension to $S^*_n(t)$

We may write

$$y^*(x,t) = (e^{tx-t^2/2} - 1 - tx)e^{-t^2/2}\ h(t)$$

where

$$h(t) = t^{-2}h_1(t)$$

and, as $t \to \infty$,

$$h(t) = 1 + O(t^2e^{-t^2})$$

and the derivatives of $h(t)$ approach zero, $h_1(t) > 0$, and $h_1(t)$ is analytic. For $t \neq 0$,

$$y^*(x,t) = [y(x,t) - e^{-t^2/2}] + [y(x,t) - e^{-t^2/2}][h(t) - 1] - xte^{-t^2/2}\ h(t).$$

Thus,

$$S^*_n(t) = S_n(t) + S_n(t)[h(t) - 1] - n^{1/2}\bar{X}te^{-t^2/2}\ h(t) \tag{41}$$

and

$$\begin{aligned}\sup_{t^2>2\log_3 n} S_n^*(t) &= \sup_{t^2>2\log_3 n} S_n(t) + O_p(\log_2 n)^{1/2}(\log_3 n)(\log_2 n)^{-2} \\ &\quad + O_p(1)(\log_3 n)^{1/2}(\log_2 n)^{-1} \\ &= \sup_t S_n(t) + o_p(\log_2 n)^{-1/2}. \end{aligned} \tag{42}$$

The covariance function of $S_n^*(t)$ is

$$\begin{aligned}\rho^*(s,t) &= [e^{st} - 1 - st]e^{-s^2/2}h(s)e^{-t^2/2}\, h(t) \\ &= e^{-(s-t)^2/2}h(s)h(t)[1-(1+st)e^{-st}].\end{aligned}$$

If s and t have the same sign,

$$\begin{aligned}E\{[S_n^*(s) - S_n^*(t)]^2\} &= 2[1-\rho^*(s,t)] \\ &\leq c(s-t)^2\end{aligned} \tag{43}$$

for some constant c. Hence the Kolmogorov bound gives

$$\sup_{t^2\leq 2\log_3 n} |S_n^*(t) - S_n^*(0)| = O_p(\log_3 n)^{1/2},$$

and with probability approaching one,

$$\begin{aligned}\sup_t S_n^*(t) &= \sup_{t^2>2\log_3 n} S_n^*(t) \\ &= \sup_t S_n(t) + o_p(\log_2 n)^{-1/2}\end{aligned} \tag{44}$$

which yields the desired result, Equation 3, for S_n^*.

REFERENCES

Billingsley, P. (1968). Convergence of probability measures. J. Wiley. New York.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statist.* **25**, 573-578.

Chernoff, H. and Lander, E. (1989). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. Technical Report. *Harvard University.*

Ghosh, J.K. and Sen, P.K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. *Proc. Berkeley Conference in honor of J. Neyman and J. Kiefer*, Vol. II, pp 789-906.

Hartigan, J.A. (1985). A failure of likelihood ratio asymptotics for normal mixtures. *Proc. Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer.* Wadsworth Advanced Books, Monterey, CA and Institute of Mathematical Statistics. Hayward, CA.

Komlos, J., Major, P., Tusnady, G. (1975). An approximation of partial sums of independent r.v.s and the sample distribution function. *Z. Wahrsch. verw. Geb.* **32**, 111-131.

Leadbetter, M.R., Lindgren, G., Rootzen, H. (1983). Extremes and related properties of random sequences. *Springer Verlag.* New York.

Department of Statistics
Statistical Laboratory
University of California, Bekeley
Berkeley, California 94720
USA

Science Center
Department of Statistics
Harvard University
1 Oxford Street
Cambridge, MA 02138, USA

The Annals of Statistics
1998, Vol. 26, No. 4, 1614–1635

ASYMPTOTIC NORMALITY OF THE MAXIMUM-LIKELIHOOD ESTIMATOR FOR GENERAL HIDDEN MARKOV MODELS

BY PETER J. BICKEL,[1] YA'ACOV RITOV[1] AND TOBIAS RYDÉN[2]

University of California, Hebrew University and Lund University

Hidden Markov models (HMMs) have during the last decade become a widespread tool for modeling sequences of dependent random variables. Inference for such models is usually based on the maximum-likelihood estimator (MLE), and consistency of the MLE for general HMMs was recently proved by Leroux. In this paper we show that under mild conditions the MLE is also asymptotically normal and prove that the observed information matrix is a consistent estimator of the Fisher information.

1. Introduction. A hidden Markov model (HMM) is a discrete-time stochastic process $\{(X_k, Y_k)\}$ such that (i) $\{X_k\}$ is a finite-state Markov chain, and (ii) given $\{X_k\}$, $\{Y_k\}$ is a sequence of conditionally independent random variables with the conditional distribution of Y_n depending on $\{X_k\}$ only through X_n. The Markov chain $\{X_k\}$ is sometimes called the *regime*. The name HMM is motivated by the assumption that $\{X_k\}$ is not observable, so that inference and so on has to be based on $\{Y_k\}$ alone. HMMs have during the last decade become widespread for modeling sequences of weakly dependent random variables, with applications in areas such as speech processing [Rabiner (1989)], neurophysiology [Fredkin and Rice (1992)] and biology [Leroux and Puterman (1992)]. See also the monograph by MacDonald and Zucchini (1997). Commonly, the conditional distributions of Y_n given X_n belong to a single parametric family, such as the normal or Poisson families, so that X_n selects the parameter used to generate Y_n. The distribution of Y_n, that is, the marginal distribution of $\{Y_k\}$, will then be a finite mixture from the parametric family. Mixtures are frequently used in i.i.d. settings to increase the dispersion governed by a specific parametric family, and this effect is obviously found in the marginal distribution of an HMM as well. In addition, $\{Y_k\}$ is dependent. HMMs can thus be viewed as an extension of Markov chains, but also as an extension of mixture models.

Inference for HMMs was first considered by Baum and Petrie, who treated the case when $\{Y_k\}$ takes values in a finite set. In Baum and Petrie (1966), results on consistency and asymptotic normality of the maximum-likelihood estimator (MLE) are given, and the conditions for consistency are weakened in Petrie (1969). In the latter paper the identifiability problem is also discussed,

Received January 1997; revised October 1997.

[1]Supported in part by NSF Grant DMS-91-15577 and by US–Israel Bi-National Science Foundation Grant 90-00031/2.

[2]Supported by the Swedish Natural Science Research Council Contract M-AA/MA 10538-303.

AMS 1991 *subject classification.* Primary 62M09.

Key words and phrases. Hidden Markov model, incomplete data, missing data, asymptotic normality.

that is, under what conditions there are no other parameters that induce the same law for $\{Y_k\}$ as the true parameter does. For general HMMs, Lindgren (1978) constructed consistent and asymptotically normal estimators of the parameters determining the conditional densities of Y_n given X_n, but he did not consider estimation of the transition probabilities. Later, Leroux (1992) proved consistency of the MLE for general HMMs under mild conditions, and local asymptotic normality (LAN) has been proved by Bickel and Ritov (1996).

The topic of the present paper is asymptotic normality of the MLE. Although Bickel and Ritov (1996) prove that an estimator similar to the MLE is asymptotically normal and achieves the information bound, their result falls short of proving that the likelihood function has a second derivative and that the MLE itself is asymptotically normal. Asymptotic normality of the MLE can be inferred from their paper, but an extra argument is needed; see Ritov (1996). In this paper we show that the curvature of the likelihood function is, asymptotically, equal to the information bound and hence the MLE is asymptotically normal. We also work with conditions that are weaker than those in Bickel and Ritov (1996).

Before we proceed, we need to introduce some notation. We let $\{X_k\}_{k=1}^{\infty}$ be a stationary Markov chain on $\{1,\dots,K\}$ with transition probabilities $\alpha(a,b) = P(X_{k+1} = b \mid X_k = a)$. We also let $\{Y_k\}$ be an $\mathscr{Y}$-valued sequence such that given $\{X_k\}$, $\{Y_k\}$ is a sequence of conditionally independent random variables, Y_n having (conditional) density $g(y|X_n)$ with respect to some σ-finite measure ν on $\mathscr{Y}$. Usually $\mathscr{Y}$ is a subset of $\mathbb{R}^q$ for some q, but it may also be a higher dimensional space. Moreover, both $\{\alpha(a,b)\}$ and $\{g(\cdot|a)\}$ depend on a parameter ϑ, that is $\alpha(a,b) = \alpha_\vartheta(a,b)$ and $g(\cdot|a) = g_\vartheta(\cdot|a)$, where ϑ is to be estimated from a realization of $\{Y_k\}$. The set to which ϑ belongs is denoted by Θ, and we assume $\Theta \subseteq \mathbb{R}^d$. Note that the stationary distribution of $\{X_k\}$, denoted by $\{\pi(a)\}_{a=1}^{K}$, does also depend on ϑ.

The most common set-up is that where ϑ contains the transition probabilities themselves, together with some parameters characterizing the g's. In particular, it is often the case that $g_\vartheta(y|a) = f(y;\phi(a))$ for some parametric family $f(y;\phi)$. We refer to this situation as the "usual parametrization." We now give a few examples of HMMs.

EXAMPLE 1 (Mixture of normal distributions). Let $K = 2$, $\vartheta = (\alpha(1,2), \alpha(2,1), \mu(1), \mu(2), \sigma^2)$ and $g_\vartheta(y|a) = \sigma^{-1}\varphi((y-\mu(a))/\sigma)$, where $\varphi(\cdot)$ is the standard normal density. Hence, $\mathscr{Y} = \mathbb{R}$ and ν is Lebesgue measure. The distribution of Y_n is a mixture of two normal distributions with different means but equal variances. This model has been used to model electric current through channels in ion membranes; see Guttorp [(1995), page 109], for a short description and Fredkin and Rice (1992) for a fuller treatment.

EXAMPLE 2 (Mixture of Poisson distributions). Let $K = 2$, $\vartheta = (\alpha(1,2), \alpha(2,1), \mu(1), \mu(2))$, and let $g_\vartheta(y|a)$ be the Poisson density with mean $\mu(a)$. Hence, $\mathscr{Y} = \{0,1,2,\dots\}$ and ν is counting measure. The distribution of Y_n is a mixture of two Poisson distributions. Albert (1991) proposed this HMM as a model for series of daily counts of epileptic seizures in one patient [see also Le, Leroux and Puterman (1992) and MacDonald and Zucchini (1997), page 146], Leroux and Puterman (1992) used it for modeling fetal lamb movements.

EXAMPLE 3 (Markov-modulated Poisson process). Let $\{X(t)\}$ be a continuous-time Markov chain on $\{1, \dots, K\}$ with intensity matrix $Q = \{q(i, j)\}$, let $\lambda(1), \dots, \lambda(K)$ be nonnegative numbers and let $\{N(t)\}$ be a doubly stochastic Poisson process (or Cox process) with random intensity function $\{\lambda(X(t))\}$; that is, given $\{\lambda(X(t))\}$, $\{N(t)\}$ has conditionally independent increments and $N(t+s) - N(t)$ has a Poisson distribution with mean $\int_t^{t+s} \lambda(X(u))\, du$. Such processes are called Markov-modulated Poisson processes, and they have been proposed for modeling traffic streams in complex telecommunication networks. See, for example, Heffes and Lucantoni (1986). The parameters of the model are the q's and the λ's. To make the connection to discrete-time HMMs, let $T_0 = 0$, let T_k be the time of the kth event in $\{N(t)\}$, $Y_k = T_k - T_{k-1}$ and $X_k = X(T_k)$. Then $\{(X_k, Y_k)\}$ is an HMM, except that given $\{X_k\}$, the distribution of Y_n depends on both X_{n-1} and X_n. Replacing $\{X_k\}$ by $\{X'_k\} = \{(X_{k-1}, X_k)\}$ takes us back to the standard set-up, however.

The joint density of $(X_1, \dots, X_n, Y_1, \dots, Y_n)$ [with respect to (counting measure)$^n \times \nu^n$] is given by

$$p_\vartheta(x_1, \dots, x_n, y_1, \dots, y_n) = \pi_\vartheta(x_1) \prod_{k=1}^{n-1} \alpha_\vartheta(x_k, x_{k+1}) \prod_{k=1}^{n} g_\vartheta(y_k | x_k),$$

and the joint density of $(Y_1, \dots, Y_n)$ (with respect to ν^n) is

$$p_\vartheta(y_1, \dots, y_n) = \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} p_\vartheta(x_1, \dots, x_n, y_1, \dots, y_n); \tag{1}$$

here, as well in the sequel, p is used as a generic symbol for densities. Looking at (1), one might think that the complexity for computing $p_\vartheta(y_1, \dots, y_n)$ is exponential in n. Fortunately, we can compute the likelihood much faster by introducing the matrix $G_\vartheta(y) = \operatorname{diag}\{g_\vartheta(y|a)\}$ and noting that

$$p_\vartheta(y_1, \dots, y_n) = \pi_\vartheta \left\{ \prod_{k=1}^{n} G_\vartheta(y_k) A_\vartheta \right\} \mathbf{1}, \tag{2}$$

where $A_\vartheta = \{\alpha_\vartheta(a, b)\}$ and $\mathbf{1}$ is a $K \times 1$-vector of ones. The computational complexity of (2) is only linear in n. A further useful observation is that conditional on the Y's, $\{X_k\}$ is still a Markov chain, although nonhomogeneous. It mixes geometrically fast, however, and this is the key to our analysis below.

The MLE, denoted by $\widehat{\vartheta}_n$, maximizes $p_\vartheta(Y_1, \dots, Y_n)$ over the parameter set Θ. In many cases we may renumber the state space of $\{X_k\}$ and the g's, leaving the likelihood unchanged, and the MLE is then not unique. In particular we may do so if the usual parametrization is employed. This ambiguity is obviously not a big concern, though.

In practice, the MLE is often computed using the EM (expectation-maximization) algorithm; $\{X_k\}$ then play the role as missing data. In the context of HMMs, the EM algorithm was formulated by Baum and co-workers; see, for example, Baum, Petrie, Soules and Weiss (1970). A recent general reference is the monograph by McLachlan and Krishnan (1997). For HMMs with the usual parametrization, the M-step, in which the parameters are updated, is always explicit in the transition probabilities; that is, the new α's are obtained without a numerical search. If the parametric family $f(y;\phi)$ is an exponential family, the M-step is often explicit in the ϕ's as well. The E-step, in which conditional expectations are evaluated, is computationally more demanding. In most cases it is carried out using the so-called forward–backward algorithm, the complexity of which is linear in n; we refer to Rabiner (1989) and Leroux and Puterman (1992) for details. The major drawback of the EM algorithm is its rate of convergence, which is only linear in the vicinity of the MLE. Various modifications of the basic algorithm have been suggested to improve on this; see, for example, Jamshidian and Jennrich (1997), Meng and van Dyk (1997) and references therein. Little has been published on which of these modifications perform well for HMMs, however.

Alternatively, one may maximize (2) with respect to ϑ directly, using any standard numerical optimization scheme. The downhill simplex algorithm [see for example Press, Flannery, Teukolsky and Vetterling (1989)], is particularly attractive since it does not require any derivatives of the objective function, and derivatives of (2) are time-consuming to compute.

Whatever optimization algorithm is used, one always faces the problem that the likelihood surface of an HMM in general is multimodal. Any algorithm, including EM, may thus converge towards a local maximum or even a saddle point. Today there are no methods guaranteed to find the MLE, but the best advice available is to start the optimization algorithm from several different, possibly random, points in Θ.

2. Further notation and assumptions. The true parameter is denoted by ϑ_0. We deliberately replace the subindex ϑ_0 by '0' in notation like P_{ϑ_0} (becoming P_0) and so on. The $\mathbb{L}_q(P_0)$-norm will be denoted $\|\cdot\|_q$; that is, $\|\cdot\|_q = \{E_0|\cdot|^q\}^{1/q}$. Sometimes $Y_m, \dots, Y_n$ will be abbreviated $\mathbf{Y}_m^n$, with an entirely similar notation for the X-process. The symbol D denotes differentiation with respect to ϑ, with D forming the gradient and D^2 forming the Hessian. Occasionally we will use a dot instead of D and two dots instead of D^2. Finally, C denotes a generic constant, finite and nonnegative, whose value may change from one expression to another.

The following assumptions will be referred to in the sequel.

(A1) The transition probability matrix $\{\alpha_0(a,b)\}$ is ergodic, that is, irreducible and aperiodic.

(A2) For all a and b, the maps $\vartheta \mapsto \alpha_\vartheta(a,b)$ and $\vartheta \mapsto \pi_\vartheta(a)$ have two continuous derivatives in some neighborhood $|\vartheta - \vartheta_0| < \delta$ of ϑ_0. For all a

and $y \in \mathscr{Y}$, the map $\vartheta \mapsto g_\vartheta(y|a)$ has two continuous derivatives in the same neighborhood.

(A3) Write $\vartheta = (\vartheta_1, \dots, \vartheta_d)$. There exists a $\delta > 0$ such that (i) for all $1 \le i \le d$ and all a,

$$E_0\left[\sup_{|\vartheta-\vartheta_0|<\delta}\left|\frac{\partial}{\partial\vartheta_i}\log g_\vartheta(Y_1|a)\right|^2\right] < \infty;$$

(ii) for all $1 \le i,\ j \le d$ and all a,

$$E_0\left[\sup_{|\vartheta-\vartheta_0|<\delta}\left|\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}\log g_\vartheta(Y_1|a)\right|\right] < \infty;$$

(iii) for $j = 1, 2$, all $1 \le i_l \le d$, $l = 1, \dots, j$, and all a,

$$\int \sup_{|\vartheta-\vartheta_0|<\delta}\left|\frac{\partial^j}{\partial\vartheta_{i_1}\cdots\partial\vartheta_{i_j}} g_\vartheta(y|a)\right| \nu(dy) < \infty.$$

(A4) There exists a $\delta > 0$ such that with

$$\rho_0(y) = \sup_{|\vartheta-\vartheta_0|<\delta}\ \max_{1\le a,\, b\le K}\frac{g_\vartheta(y|a)}{g_\vartheta(y|b)},$$

$P_0(\rho_0(Y_1) = \infty \mid X_1 = a) < 1$ for all a.

(A5) ϑ_0 is an interior point of Θ.

(A6) The maximum-likelihood estimator is strongly consistent.

Without loss of generality, we assume that the δ's in (A2)–(A4) agree.

REMARK. If (A1) holds, $\{X_k\}$ is ergodic under P_0. This implies that $\{Y_k\}$ is ergodic as well; see Leroux [(1992), page 130]. (A2) and (A3) are essentially regularity conditions of "Cramér type," that we cannot expect to weaken considerably. (A4) fails to hold if there are two g_0's with disjoint support; let, for example, the g's be location shifts of Beta densities. Heuristically, the result is a gain of information, however, rather than a loss, and it is possible that our results could be refined to include also this case.

In (A6) we assume that $\widehat{\vartheta}_n \to \vartheta_0$, P_0-a.s. as $n \to \infty$ (up to a possible permutation of states). Consistency of the MLE is discussed by Leroux (1992), and the conditions needed to ensure (A6) are essentially the following: (i) (A1); (ii) for all a and b, the map $\vartheta \mapsto \alpha_\vartheta(a, b)$ is continuous on Θ; (iii) for all a and $y \in \mathscr{Y}$, the map $\vartheta \mapsto g_\vartheta(y|a)$ is continuous on Θ; (iv) Θ is compact (this assumption can be relaxed somewhat; see Leroux's paper); (v) for each a, $E_0|\log g_0(Y_1|a)| < \infty$; (vi) For each a and ϑ there is a $\delta > 0$ such that $E_0[\sup_{|\vartheta'-\vartheta|<\delta}(\log g_{\vartheta'}(Y_1|a))^+] < \infty$; (vii) for each ϑ such that the laws P_ϑ and P_0 agree, $\vartheta = \vartheta_0$ (up to a possible permutation of states).

Obviously, conditions (ii), (iii) and (vi) are global, whereas conditions (A2)–(A4) are all local. Condition (vii) holds, for example, if the HMM has the usual parametrization, finite mixtures of the parametric family $\{f(y;\phi)\}$ are identifiable and the ϕ_0's are distinct. Families of which finite mixtures are identifiable include the normal distribution, the Poisson distribution and the exponential distribution.

EXAMPLE 1 (Continued). We may define Θ by $\alpha(1,2), \alpha(2,1) \in [0,1]$, $\mu(a) \in [-1/\varepsilon, 1/\varepsilon]$, and $\sigma^2 \in [\varepsilon, 1/\varepsilon]$ for some small $\varepsilon > 0$. Conditions (A2)–(A4) are then all satisfied, as are the conditions for consistency listed above provided $\alpha_0(1,2), \alpha_0(2,1) \in (0,1)$ [implying (A1)].

EXAMPLE 2 (Continued). We define Θ by $\alpha(1,2), \alpha(2,1) \in [0,1]$ and $\mu(a) \in [0, 1/\varepsilon]$ for some small $\varepsilon > 0$. Then (A2)–(A4) and the consistency conditions are satisfied provided (A1) also holds.

EXAMPLE 3 (Continued). Define Θ by Q having off-diagonal elements bounded by $1/\varepsilon$ and $\lambda(a) \in [0, 1/\varepsilon]$ for some small $\varepsilon > 0$. Then (A2)–(A4) and the consistency conditions are satisfied provided (A1) also holds; it does if Q_0 is irreducible and all $\lambda_0(a) > 0$. Parameter estimation and consistency of the MLE are further discussed in Rydén (1994).

3. Main results. To prove asymptotic normality of the MLE, we need two lemmas which themselves are of considerable interest. These lemmas involve the loglikelihood, denoted by $L_n(\vartheta) = \log p_\vartheta(Y_1, \dots, Y_n)$, and the Fisher information matrix for $\{Y_k\}$, denoted by $\mathscr{I}_0$. Intuitively, $\mathscr{I}_0$ may be thought of as the limiting covariance matrix of either $n^{-1/2}\dot{L}_n(\vartheta_0)$ or $D \log p_{\vartheta_0}(Y_n | Y_{n-1}, \dots, Y_1)$. In Section 4 we show that both of these definitions are valid.

The first lemma is a central limit theorem for the score function at ϑ_0.

LEMMA 1. *Assume that* (A1)–(A4) *hold. Then* $n^{-1/2}\dot{L}_n(\vartheta_0) \to \mathscr{N}(0, \mathscr{I}_0)$ P_0-*weakly as* $n \to \infty$.

We prove this lemma in Section 4. The second lemma is a law of large numbers for the Hessian of the log likelihood.

LEMMA 2. *Assume that* (A1)–(A4) *hold and let* ϑ_n^* *be any, possibly stochastic, sequence in* Θ *such that* $\vartheta_n^* \to \vartheta_0$, P_0-*a.s. as* $n \to \infty$. *Then* $n^{-1}\ddot{L}_n(\vartheta_n^*) \to -\mathscr{I}_0$ *in* P_0-*probability as* $n \to \infty$.

This result will be proved in Section 5. Note that Lemma 2 shows that if (A1)–(A4) and (A6) hold, the observed information, that is $-n^{-1}\ddot{L}_n(\widehat{\vartheta}_n)$, converges to $\mathscr{I}_0$ in P_0-probability. The main result is now as follows.

THEOREM 1. *Assume that* (A1)–(A6) *hold and that* $\mathscr{I}_0$ *is nonsingular. Then* $n^{1/2}(\widehat{\vartheta}_n - \vartheta_0) \to \mathscr{N}(0, \mathscr{I}_0^{-1})$, P_0-*weakly as* $n \to \infty$.

PROOF. The proof essentially uses the approach introduced by Cramér. For n large enough, $\widehat{\vartheta}_n$ is an interior point of Θ and $|\widehat{\vartheta}_n - \vartheta_0| < \delta$, and we can then make a Taylor expansion of $\dot{L}_n$ about ϑ_0,

$$0 = \dot{L}_n(\widehat{\vartheta}_n) = \dot{L}_n(\vartheta_0) + \ddot{L}(\overline{\vartheta}_n)(\widehat{\vartheta}_n - \vartheta_0),$$

where $\overline{\vartheta}_n$ is a point on the line segment between ϑ_0 and $\widehat{\vartheta}_n$. Rewriting this expression, we obtain

$$n^{1/2}(\widehat{\vartheta}_n - \vartheta_0) = \left[-n^{-1}\ddot{L}_n(\overline{\vartheta}_n)\right]^{-1} n^{-1/2}\dot{L}_n(\vartheta_0).$$

The result now follows from the above lemmas. □

REMARK. Lemmas 1 and 2 also imply LAN of our model. In fact, they even imply uniform LAN, that is, that in the expansion

$$L_n(\vartheta_0 + n^{-1/2}u) - L_n(\vartheta_0) = n^{-1/2}u^T\dot{L}_n(\vartheta_0) + n^{-1}\tfrac{1}{2}u^T\ddot{L}_n(\vartheta_0)u + R_n(u),$$

$R_n(u)$ tends to zero in P_0-probability uniformly over compact subsets of $\mathbb{R}^d$. The superindex T denotes transpose.

Throughout the remainder of the paper, we shall make two assumptions that simplify the notation but do not remove any principal difficulties. The first assumption is that ϑ is one-dimensional, which saves us from using notation like uu^T. At one instance we do use this notation, namely, in the definition of the Fisher information matrix below. Our second assumption concerns the transition probabilities. By (A1), there exists a positive integer r such that all r-step transition probabilities $\alpha_0^{(r)}(a,b) = P_0(X_r = b \mid X_0 = a) > 0$. The assumption we make is that this inequality is satisfied with $r = 1$. We comment on the general case after Lemma 3.

4. A central limit theorem for the score function. Since the bivariate process $\{(X_k, Y_k)\}$ is stationary, we may extend it to a doubly infinite stationary sequence $\{(X_k, Y_k)\}_{k=-\infty}^{\infty}$, a feature that we will use frequently. Let $p_\vartheta(Y_1|Y_0, \ldots, Y_{-n})$ denote the conditional density of Y_1 given $Y_0, \ldots, Y_{-n}$. By the very definition of an HMM,

$$p_\vartheta(Y_1|\mathbf{Y}_{-n}^0) = \sum_{a=1}^{K} g_\vartheta(Y_1|a)P_\vartheta(X_1 = a \mid \mathbf{Y}_{-n}^0). \tag{3}$$

By a martingale convergence theorem by Lévy [see, e.g., Shiryayev (1984), page 478], $P_\vartheta(X_1 = a \mid \mathbf{Y}_{-n}^0) \to P_\vartheta(X_1 = a \mid \mathbf{Y}_{-\infty}^0)$ P_ϑ-a.s. as $n \to \infty$. Thus, if we define $p_\vartheta(Y_1|Y_0, Y_{-1}, \ldots)$ in analogy with (3), $p_\vartheta(Y_1|\mathbf{Y}_{-n}^0) \to p_\vartheta(Y_1|\mathbf{Y}_{-\infty}^0)$ P_ϑ-a.s.

Now, by a general identity for models with missing data [see Louis (1982), page 227], valid in our case because the X's take values in a finite set,

$$
\begin{aligned}
&D\log p_\vartheta(Y_1|Y_0,\dots,Y_{-n})\\
&\quad= D\log p_\vartheta(Y_{-n},\dots,Y_1) - D\log p_\vartheta(Y_{-n},\dots,Y_0)\\
&\quad= E_\vartheta[D\log p_\vartheta(X_{-n},\dots,X_1,Y_{-n},\dots,Y_1)\mid Y_{-n},\dots,Y_1]\\
&\qquad - E_\vartheta[D\log p_\vartheta(X_{-n},\dots,X_1,Y_{-n},\dots,Y_0)\mid Y_{-n},\dots,Y_0];
\end{aligned}
\tag{4}
$$

note that in the second term on the right-hand side, we consider X_1 as missing despite that Y_1 is not observed, a trick that will simplify the following computations slightly. Thus, writing $\lambda_\vartheta(a,b) = D\log\alpha_\vartheta(a,b)$, $\gamma_\vartheta(y|a) = D\log g_\vartheta(y|a)$, and $\tau_\vartheta(a) = D\log\pi_\vartheta(a)$, we have

$$
\begin{aligned}
&D\log p_{\vartheta_0}(Y_1|Y_0,\dots,Y_{-n})\\
&\quad= \sum_{k=-n}^{0}\Big\{E_0[\gamma_0(Y_k|X_k)+\lambda_0(X_k,X_{k+1})\mid \mathbf{Y}_{-n}^1]\\
&\qquad\qquad - E_0[\gamma_0(Y_k|X_k)+\lambda_0(X_k,X_{k+1})\mid \mathbf{Y}_{-n}^0]\Big\}\\
&\qquad + E_0[\gamma_0(Y_1|X_1)\mid \mathbf{Y}_{-n}^1] + E_0[\tau_0(X_{-n})\mid \mathbf{Y}_{-n}^1] - E_0[\tau_0(X_{-n})\mid \mathbf{Y}_{-n}^0].
\end{aligned}
\tag{5}
$$

Define

$$
\begin{aligned}
\eta_1 = &\sum_{k=-\infty}^{0}\Big\{E_0[\gamma_0(Y_k|X_k)+\lambda_0(X_k,X_{k+1})\mid \mathbf{Y}_{-\infty}^1]\\
&\qquad - E_0[\gamma_0(Y_k|X_k)+\lambda_0(X_k,X_{k+1})\mid \mathbf{Y}_{-\infty}^0]\Big\}\\
&+ E_0[\gamma_0(Y_1|X_1)\mid \mathbf{Y}_{-\infty}^1].
\end{aligned}
\tag{6}
$$

The sum in (6) is absolutely convergent in $\mathbb{L}_2(P_0)$, so that the right-hand side of (6) defines a random variable in $\mathbb{L}_2(P_0)$. We do not show this here, but it follows from the proof of Lemma 6 below. Under somewhat stronger conditions, the result $\eta_1 \in \mathbb{L}_2(P_0)$ is shown in Lemma 2.3 in Bickel and Ritov (1996). We now define the Fisher information matrix as $\mathscr{I}_0 = E_0[\eta_1\eta_1^T]$. Before proving Lemma 1, we give some additional notation and lemmas.

Note that if (A1) and (A2) hold, there exist a $\delta > 0$ and a $\sigma_0 > 0$ such that $\inf\{\alpha_\vartheta(a,b)\colon a,b,|\vartheta-\vartheta_0|<\delta\} \ge \sigma_0$, $\inf\{\alpha_\vartheta^*(a,b)\colon a,b,|\vartheta-\vartheta_0|<\delta\} \ge \sigma_0$ and $\inf\{\pi_\vartheta(a)\colon a,|\vartheta-\vartheta_0|<\delta\} \ge \sigma_0$, where $\alpha_\vartheta^*(a,b) = \pi_\vartheta(b)/\pi_\vartheta(a)\times\alpha_\vartheta(b,a)$ are the transition probabilities of the time-reversed version of $\{X_k\}$ (recall that we assume $r = 1$). Without loss of generality, we assume that this δ agrees with the one in (A2)–(A4). Let

$$\mu_0(y) = \{1 + (K-1)\sigma_0^{-2}\rho_0(y)\}^{-1};$$

if (A4) holds, $P_0(\mu_0(Y_1) > 0 \mid X_1 = a) > 0$ for all a. For further reference, we cite the following result from Bickel and Ritov (1996); it is their Lemma 3.3.

LEMMA 3. *Let $-n \le l < k \le 0$ and let H_k be an event defined in terms of $X_k, X_{k+1}, \dots, X_0$ and $Y_k, Y_{k+1}, \dots, Y_0$ only. Then for all ϑ such that $|\vartheta - \vartheta_0| < \delta$,*

$$\begin{aligned}
\max_a P_\vartheta(H_k \mid \mathbf{Y}_{-n}^0, X_l = a) &- \min_a P_\vartheta(H_k \mid \mathbf{Y}_{-n}^0, X_l = a) \\
&\le \prod_{i=l+1}^{k-1} (1 - 2\mu_0(Y_i)) \\
&\le \prod_{i=l+1}^{k-1} \exp(-2\mu_0(Y_i)).
\end{aligned}$$

REMARK. If $r > 1$, the result corresponding to Lemma 3 (and with an entirely similar proof) reads

$$\begin{aligned}
&\max_a P_\vartheta(H_k \mid \mathbf{Y}_{-n}^0, X_{k-qr} = a) - \min_a P_\vartheta(H_k \mid \mathbf{Y}_{-n}^0, X_{k-qr} = a) \\
&\qquad \le \prod_{i=2}^{q} \exp(-2\mu_0(Y_{k-ir+1}, \dots, Y_{k-ir+2r-1})),
\end{aligned} \tag{7}$$

where now

$$\mu_0(y_1, \dots, y_{2r-1}) = \frac{1}{1 + (K-1)\sigma_0^{-2} \prod_{i=1}^{2r-1} \rho(y_i)},$$

and with σ_0 defined as above but in terms of the r-step transition probabilities. By deleting every second factor in (7) we obtain a bound with factors containing disjoint blocks of Y's. The proofs below then go through as when $r = 1$, except for some very minor changes caused by the need to work with the Y's in blocks of size r.

LEMMA 4. *Let $-n \le k \le 0$ and define*

$$S_\vartheta(n, k) = \max_{a,b,c} \big| P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^0,\ X_1 = b) - P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^0,\ X_1 = c) \big|.$$

Then, for any ϑ such that $|\vartheta - \vartheta_0| < \delta$,

$$S_\vartheta(n, k) \le \prod_{i=k+1}^{0} \exp(-2\mu_0(Y_i)).$$

The proof follows from Lemma 3 and the observation that the time-reversed version of $\{(X_k, Y_k)\}$ is an HMM as well.

$$\leq \max_{b,c}\left|P_\vartheta(X_k = a \mid X_{-n} = b, \mathbf{Y}_{-n+1}^1) - P_\vartheta(X_k = a \mid X_{-n} = c, \mathbf{Y}_{-n+1}^1)\right|$$

$$\leq \prod_{i=-n+1}^{k-1} \exp(-2\mu_0(Y_i)),$$

the third part holds; the last inequality follows from Lemma 3. When $\mathbf{Y}_{-n}^1$ and $\mathbf{Y}_{-m}^1$ are replaced by $\mathbf{Y}_{-n}^0$ and $\mathbf{Y}_{-m}^0$, respectively, the bound follows in a completely similar fashion.

The last part is proved using part three and an argument like the one used to prove part two. Finally, if n or m is infinite, use the fact that $P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^1) \to P_\vartheta(X_k = a \mid \mathbf{Y}_{-\infty}^1)$ P_ϑ-a.s. and so on. □

We are now ready to prove the following result.

LEMMA 6. *There exist constants* $\beta_0 \in [0, 1)$ *and* C_0 *such that*

$$\|D \log p_{\vartheta_0}(Y_1|Y_0, \ldots, Y_{-n}) - \eta_1\|_2 \leq C_0\beta_0^n.$$

PROOF. Comparing (5) and (6), we see that it is sufficient to prove that there are $\beta_0 \in [0, 1)$ and C_0 such that

$$\left|\left|E_0[\tau_0(X_{-n}) \mid \mathbf{Y}_{-n}^1] - E_0[\tau_0(X_{-n}) \mid \mathbf{Y}_{-n}^0]\right|\right|_2 \leq C_0\beta_0^n, \tag{8}$$

$$\left|\left|E_0[\gamma_0(Y_1|X_1) \mid \mathbf{Y}_{-n}^1] - E_0[\gamma_0(Y_1|X_1) \mid \mathbf{Y}_{-\infty}^1]\right|\right|_2 \leq C_0\beta_0^n, \tag{9}$$

$$\left|\left|\sum_{k=-\lfloor n/2\rfloor}^{0} \{E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-n}^j] - E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-\infty}^j]\}\right|\right|_2 \leq C_0\beta_0^n, \tag{10}$$

$$\left|\left|\sum_{k=-\lfloor n/2\rfloor}^{0} \{E_0[\lambda_0(X_k, X_{k+1}) \mid \mathbf{Y}_{-n}^j] - E_0[\lambda_0(X_k, X_{k+1}) \mid \mathbf{Y}_{-\infty}^j]\}\right|\right|_2 \leq C_0\beta_0^n, \tag{11}$$

$$\left|\left|\sum_{k=-n}^{-\lfloor n/2\rfloor-1} \{E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-n}^1] - E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-n}^0]\}\right|\right|_2 \leq C_0\beta_0^n, \tag{12}$$

$$\left|\left|\sum_{k=-n}^{-\lfloor n/2\rfloor-1} \{E_0[\lambda_0(X_k, X_{k+1}) \mid \mathbf{Y}_{-n}^1] - E_0[\lambda_0(X_k, X_{k+1}) \mid \mathbf{Y}_{-n}^0]\}\right|\right|_2 \leq C_0\beta_0^n, \tag{13}$$

$$\left|\left|\sum_{k=-\infty}^{-\lfloor n/2\rfloor-1} \{E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-\infty}^1] - E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-\infty}^0]\}\right|\right|_2 \leq C_0\beta_0^n, \tag{14}$$

$$\left|\left|\sum_{k=-\infty}^{-\lfloor n/2\rfloor-1} \{E_0[\lambda_0(X_k, X_{k+1}) \mid \mathbf{Y}_{-\infty}^1] - E_0[\lambda_0(X_k, X_{k+1}) \mid \mathbf{Y}_{-\infty}^0]\}\right|\right|_2 \leq C_0\beta_0^n \tag{15}$$

for $j = 0, 1$, where $\lfloor\cdot\rfloor$ denotes the integer part.

We start with (8). By the first part of Lemma 5 we have

$$\begin{aligned}
&\left|E_0[\tau_0(X_{-n}) \mid \mathbf{Y}_{-n}^1] - E_0[\tau_0(X_{-n}) \mid \mathbf{Y}_{-n}^0]\right| \\
&\quad = \left|\sum_{a=1}^{K} \tau_0(a)\,[P_0(X_{-n} = a \mid \mathbf{Y}_{-n}^1) - P_0(X_{-n} = a \mid \mathbf{Y}_{-n}^0)]\right| \\
&\quad \le \max_a \tau_0(a)\, C \prod_{i=-n+1}^{0} \exp(-2\mu_0(Y_i)).
\end{aligned}$$

Thus, by the definition of an HMM,

$$\begin{aligned}
&\left\|E_0[\tau_0(X_{-n}) \mid \mathbf{Y}_{-n}^1] - E_0[\tau_0(X_{-n}) \mid \mathbf{Y}_{-n}^0]\right\|_2^2 \\
&\quad \le CE_0\left[\prod_{i=-n+1}^{0} \exp(-4\mu_0(Y_i))\right] \\
&\quad = CE_0\left[E_0\left[\prod_{i=-n+1}^{0} \exp(-4\mu_0(Y_i)) \,\middle|\, \mathbf{X}_{-n+1}^0\right]\right] \\
&\quad = CE_0\left[\prod_{i=-n+1}^{0} E_0\big[\exp(-4\mu_0(Y_i)) \,\big|\, X_i\big]\right] \\
&\quad \le CE_0\left[\prod_{i=-n+1}^{0} \max_a E_0\big[\exp(-4\mu_0(Y_i)) \,\big|\, X_i = a\big]\right] \\
&\quad = C\beta^n
\end{aligned}$$

for some $\beta \in [0, 1)$ and (8) follows. A similar argument shows (9).

We now turn to (10). By the third part of Lemma 5, with $m = \infty$,

$$\begin{aligned}
&\left|E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-n}^j] - E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-\infty}^j]\right| \\
&\quad = \left|\sum_{a=1}^{K} \gamma_0(Y_k|a)\,[P_0(X_k = a \mid \mathbf{Y}_{-n}^j) - P_0(X_k = a \mid \mathbf{Y}_{-\infty}^j)]\right| \\
&\quad \le \max_a |\gamma_0(Y_k|a)|\, C \prod_{i=-n+1}^{k-1} \exp(-2\mu_0(Y_i))
\end{aligned}$$

P_0-a.s. Thus,

$$\begin{aligned}
&\left\|E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-n}^j] - E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-\infty}^j]\right\|_2^2 \\
&\quad \le E_0\left[C \max_a |\gamma_0(Y_k|a)|^2 \prod_{i=-n+1}^{k-1} \exp(-4\mu_0(Y_i))\right] \\
&\quad \le CE_0\left[E_0\left[\max_a |\gamma_0(Y_k|a)|^2 \prod_{i=-n+1}^{k-1} \exp(-4\mu_0(Y_i)) \,\middle|\, \mathbf{X}_{-n+1}^k\right]\right]
\end{aligned}$$

$$= CE_0\Big[E_0\Big[\max_a |\gamma_0(Y_k|a)|^2 \mid X_k\Big] \prod_{i=-n+1}^{k-1} E_0\big[\exp(-4\mu_0(Y_i)) \mid X_i\big]\Big]$$

$$\le C \max_b E_0\Big[\max_a |\gamma_0(Y_k|a)|^2 \mid X_k = b\Big]\beta^{k-1+n},$$

so that

$$\Bigg\| \sum_{k=-\lfloor n/2\rfloor}^{0} \{E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-n}^{j}] - E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-\infty}^{j}]\}\Bigg\|_2$$

$$\le C \sum_{k=-\lfloor n/2\rfloor}^{0} \beta^{(k-1+n)/2} \le C\beta^{(-\lfloor n/2\rfloor-1+n)/2},$$

and (10) follows. Also (11)–(15) follow in an entirely similar fashion, using other parts of Lemma 5. Note that (14) and (15) show that $\eta_1 \in \mathbb{L}_2(P_0)$. □

PROOF OF LEMMA 1. Let $\xi_k = D\log p_{\vartheta_0}(Y_k|Y_{k-1},\dots,Y_1)$, so that $\dot{L}_n(\vartheta_0) = \sum_{k=1}^n \xi_k$, and let

$$\eta_k = \sum_{i=-\infty}^{k-1} \Big\{E_0\big[\gamma_0(Y_i|X_i) + \lambda_0(X_i, X_{i+1}) \mid \mathbf{Y}_{-\infty}^{k}\big]$$
$$- E_0\big[\gamma_0(Y_i|X_i) + \lambda_0(X_i, X_{i+1}) \mid \mathbf{Y}_{-\infty}^{k-1}\big]\Big\}$$
$$+ E_0\big[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-\infty}^{k}\big].$$

Using (A3)(iii), it readily follows that

$$E_0\big[\gamma_0(Y_1|X_1) \mid \mathbf{Y}_{-\infty}^{0}\big] = E_0\big[E_0[\gamma_0(Y_1|X_1) \mid \mathbf{Y}_{-\infty}^{0}, X_1] \mid \mathbf{Y}_{-\infty}^{0}\big]$$
$$= E_0\big[E_0[\gamma_0(Y_1|X_1) \mid X_1] \mid \mathbf{Y}_{-\infty}^{0}\big] = 0,$$

so that $\{\eta_k\}$ is a stationary and ergodic (because $\{Y_k\}$ is ergodic) martingale increment sequence with respect to $\{\sigma(Y_{-\infty}^k)\}$ in $\mathbb{L}_2(P_0)$. Its covariance matrix is $\mathscr{I}_0$. By the central limit theorem for martingales [see, e.g., Durrett (1991), page 375], we obtain

$$n^{-1/2}\sum_{k=1}^{n} \eta_k \to \mathscr{N}(0, \mathscr{I}_0). \tag{16}$$

Finally, Lemma 6 shows that

$$\Bigg\| n^{-1/2}\sum_{k=1}^{n} \xi_k - n^{-1/2}\sum_{k=1}^{n} \eta_k \Bigg\|_2 \le n^{-1/2}\sum_{k=1}^{n} \|\xi_k - \eta_k\|_2$$
$$= n^{-1/2}\sum_{k=1}^{n} \|D\log p_{\vartheta_0}(Y_1|Y_0,\dots,Y_{-k+2}) - \eta_1\|_2,$$

where the last equality follows by stationarity. By Lemma 6, the expression on the right-hand side tends to zero as $n \to \infty$, whence the result follows from (16). □

5. A law of large numbers for the observed information. In this section we prove Lemma 2 via a uniform law of large numbers for the Hessian of the loglikelihood. Our approach is similar to the one used in Section 4, but the derivation is more delicate. First, again by a general identity for models with missing data [see Louis (1982), page 227], valid in our case because the X's take values in a finite set,

$$\begin{aligned}
&D^2 \log p_\vartheta(Y_1 | Y_0, \dots, Y_{-n})\\
&\quad = D^2 \log p_\vartheta(Y_{-n}, \dots, Y_1) - D^2 \log p_\vartheta(Y_{-n}, \dots, Y_0)\\
&\quad = E_\vartheta\big[D^2 \log p_\vartheta(X_{-n}, \dots, X_1, Y_{-n}, \dots, Y_1) \mid \mathbf{Y}_{-n}^1\big]\\
&\qquad + E_\vartheta\big[(D \log p_\vartheta(X_{-n}, \dots, X_1, Y_{-n}, \dots, Y_1))^2 \mid \mathbf{Y}_{-n}^1\big]\\
&\qquad - \big\{E_\vartheta\big[D \log p_\vartheta(X_{-n}, \dots, X_1, Y_{-n}, \dots, Y_1) \mid \mathbf{Y}_{-n}^1\big]\big\}^2\\
&\qquad - E_\vartheta\big[D^2 \log p_\vartheta(X_{-n}, \dots, X_1, Y_{-n}, \dots, Y_0) \mid \mathbf{Y}_{-n}^0\big]\\
&\qquad - E_\vartheta\big[(D \log p_\vartheta(X_{-n}, \dots, X_1, Y_{-n}, \dots, Y_0))^2 \mid \mathbf{Y}_{-n}^0\big]\\
&\qquad + \big\{E_\vartheta\big[D \log p_\vartheta(X_{-n}, \dots, X_1, Y_{-n}, \dots, Y_0) \mid \mathbf{Y}_{-n}^0\big]\big\}^2\\
&\quad = \sum_{k=-n}^{0} \big\{E_\vartheta\big[\dot\gamma_\vartheta(Y_k|X_k) + \dot\lambda_\vartheta(X_k, X_{k+1}) \mid \mathbf{Y}_{-n}^1\big]\\
&\qquad\qquad - E_\vartheta\big[\dot\gamma_\vartheta(Y_k|X_k) + \dot\lambda_\vartheta(X_k, X_{k+1}) \mid \mathbf{Y}_{-n}^0\big]\big\}\\
&\qquad + E_\vartheta\big[\dot\gamma_\vartheta(Y_1|X_1) \mid \mathbf{Y}_{-n}^1\big] + E_\vartheta\big[\dot\tau_\vartheta(X_{-n}) \mid \mathbf{Y}_{-n}^1\big] - E_\vartheta\big[\dot\tau_\vartheta(X_{-n}) \mid \mathbf{Y}_{-n}^0\big]\\
&\qquad + \sum_{k=-n}^{0} \sum_{l=-n}^{0} \Big\{E_\vartheta\big[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^1\big]\\
&\qquad\qquad - E_\vartheta\big[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-n}^1\big] E_\vartheta\big[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^1\big]\\
&\qquad\qquad - E_\vartheta\big[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^0\big]\\
&\qquad\qquad + E_\vartheta\big[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-n}^0\big] E_\vartheta\big[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^0\big]\\
&\qquad\qquad + E_\vartheta\big[\lambda_\vartheta(X_k, X_{k+1})\lambda_\vartheta(X_l, X_{l+1}) \mid \mathbf{Y}_{-n}^1\big]\\
&\qquad\qquad - E_\vartheta\big[\lambda_\vartheta(X_k, X_{k+1}) \mid \mathbf{Y}_{-n}^1\big] E_\vartheta\big[\lambda_\vartheta(X_l, X_{l+1}) \mid \mathbf{Y}_{-n}^1\big]\\
&\qquad\qquad - E_\vartheta\big[\lambda_\vartheta(X_k, X_{k+1})\lambda_\vartheta(X_l, X_{l+1}) \mid \mathbf{Y}_{-n}^0\big]\\
&\qquad\qquad + E_\vartheta\big[\lambda_\vartheta(X_k, X_{k+1}) \mid \mathbf{Y}_{-n}^0\big] E_\vartheta\big[\lambda_\vartheta(X_l, X_{l+1}) \mid \mathbf{Y}_{-n}^0\big]\\
&\qquad\qquad + 2E_\vartheta\big[\gamma_\vartheta(Y_k|X_k)\lambda_\vartheta(X_l, X_{l+1}) \mid \mathbf{Y}_{-n}^1\big]\\
&\qquad\qquad - 2E_\vartheta\big[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-n}^1\big] E_\vartheta\big[\lambda_\vartheta(X_l, X_{l+1}) \mid \mathbf{Y}_{-n}^1\big]\\
&\qquad\qquad - 2E_\vartheta\big[\gamma_\vartheta(Y_k|X_k)\lambda_\vartheta(X_l, X_{l+1}) \mid \mathbf{Y}_{-n}^0\big]\\
&\qquad\qquad + 2E_\vartheta\big[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-n}^0\big] E_\vartheta\big[\lambda_\vartheta(X_l, X_{l+1}) \mid \mathbf{Y}_{-n}^0\big]\Big\}\\
&\qquad + E_\vartheta\big[\gamma_\vartheta^2(Y_1|X_1) \mid \mathbf{Y}_{-n}^1\big] - \big\{E_\vartheta\big[\gamma_\vartheta(Y_1|X_1) \mid \mathbf{Y}_{-n}^1\big]\big\}^2
\end{aligned} \tag{17}$$

$$
\begin{aligned}
&+\sum_{k=-n}^{0}\Big\{2E_\vartheta\big[\gamma_\vartheta(Y_1|X_1)\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^{1}\big]\\
&\qquad\qquad -2E_\vartheta\big[\gamma_\vartheta(Y_1|X_1)\mid \mathbf{Y}_{-n}^{1}\big]E_\vartheta\big[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^{1}\big]\\
&\qquad\qquad +2E_\vartheta\big[\gamma_\vartheta(Y_1|X_1)\lambda_\vartheta(X_k, X_{k+1})\mid \mathbf{Y}_{-n}^{1}\big]\\
&\qquad\qquad -2E_\vartheta\big[\gamma_\vartheta(Y_1|X_1)\mid \mathbf{Y}_{-n}^{1}\big]E_\vartheta\big[\lambda_\vartheta(X_k, X_{k+1})\mid \mathbf{Y}_{-n}^{1}\big]\Big\}\\
&+E_\vartheta\big[\tau_\vartheta^2(X_{-n})\mid \mathbf{Y}_{-n}^{1}\big]-\big\{E_\vartheta\big[\tau_\vartheta(X_{-n})\mid \mathbf{Y}_{-n}^{1}\big]\big\}^2\\
&-E_\vartheta\big[\tau_\vartheta^2(X_{-n})\mid \mathbf{Y}_{-n}^{0}\big]+\big\{E_\vartheta\big[\tau_\vartheta(X_{-n})\mid \mathbf{Y}_{-n}^{0}\big]\big\}^2\\
&+\sum_{k=-n}^{0}\Big\{2E_\vartheta\big[\tau_\vartheta(X_{-n})\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^{1}\big]\\
&\qquad\qquad -2E_\vartheta\big[\tau_\vartheta(X_{-n})\mid \mathbf{Y}_{-n}^{1}\big]E_\vartheta\big[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^{1}\big]\\
&\qquad\qquad -2E_\vartheta\big[\tau_\vartheta(X_{-n})\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^{0}\big]\\
&\qquad\qquad +2E_\vartheta\big[\tau_\vartheta(X_{-n})\mid \mathbf{Y}_{-n}^{0}\big]E_\vartheta\big[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^{0}\big]\\
&\qquad\qquad +2E_\vartheta\big[\tau_\vartheta(X_{-n})\lambda_\vartheta(X_k, X_{k+1})\mid \mathbf{Y}_{-n}^{1}\big]\\
&\qquad\qquad -2E_\vartheta\big[\tau_\vartheta(X_{-n})\mid \mathbf{Y}_{-n}^{1}\big]E_\vartheta\big[\lambda_\vartheta(X_k, X_{k+1})\mid \mathbf{Y}_{-n}^{1}\big]\\
&\qquad\qquad -2E_\vartheta\big[\tau_\vartheta(X_{-n})\lambda_\vartheta(X_k, X_{k+1})\mid \mathbf{Y}_{-n}^{0}\big]\\
&\qquad\qquad +2E_\vartheta\big[\tau_\vartheta(X_{-n})\mid \mathbf{Y}_{-n}^{0}\big]E_\vartheta\big[\lambda_\vartheta(X_k, X_{k+1})\mid \mathbf{Y}_{-n}^{0}\big]\Big\}\\
&+2E_\vartheta\big[\tau_\vartheta(X_{-n})\gamma_\vartheta(Y_1|X_1)\mid \mathbf{Y}_{-n}^{1}\big]\\
&-2E_\vartheta\big[\tau_\vartheta(X_{-n})\mid \mathbf{Y}_{-n}^{1}\big]E_\vartheta\big[\gamma_\vartheta(Y_1|X_1)\mid \mathbf{Y}_{-n}^{1}\big].
\end{aligned}
$$

Again, we need some additional lemmas before we look closer at this expression.

LEMMA 7. *Let $-m \le -n \le k$, $l \le 0$. Then for any ϑ such that $|\vartheta - \vartheta_0| < \delta$,*

$$
\begin{aligned}
&\max_{a,b}\big|P_\vartheta(X_k = a,\ X_l = b\mid \mathbf{Y}_{-n}^{1}) - P_\vartheta(X_k = a,\ X_l = b\mid \mathbf{Y}_{-n}^{0})\big|\\
&\qquad \le \prod_{i=k\vee l+1}^{0}\exp(-2\mu_0(Y_i)),\\
&\max_{a,b}\big|P_\vartheta(X_k = a,\ X_l = b\mid \mathbf{Y}_{-n}^{1}) - P_\vartheta(X_k = a,\ X_l = b\mid \mathbf{Y}_{-m}^{1})\big|\\
&\qquad \le \prod_{i=-n+1}^{k\wedge l-1}\exp(-2\mu_0(Y_i)).
\end{aligned}
$$

The second conclusion holds true also if $\mathbf{Y}_{-n}^{1}$ and $\mathbf{Y}_{-m}^{1}$ are replaced by $\mathbf{Y}_{-n}^{0}$ and $\mathbf{Y}_{-m}^{0}$, respectively.

The proof is entirely similar to the proofs of parts two and four of Lemma 5.

LEMMA 8. *Let $-n \le k,\ l \le 0$. Then for any ϑ such that $|\vartheta - \vartheta_0| < \delta$,*

$$\max_{a,b} \left| P_\vartheta(X_k = a,\ X_l = b \mid \mathbf{Y}_{-n}^1) - P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^1) P_\vartheta(X_l = b \mid \mathbf{Y}_{-n}^1) \right| \le \prod_{i=k\wedge l+1}^{k\vee l-1} \exp(-2\mu_0(Y_i)).$$

The conclusion holds true also if $\mathbf{Y}_{-n}^1$ is replaced by $\mathbf{Y}_{-n}^0$.

PROOF. Assume that $k \ge l$. Then

$$\begin{aligned}
&\left| P_\vartheta(X_k = a,\ X_l = b \mid \mathbf{Y}_{-n}^1) - P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^1) P_\vartheta(X_l = b \mid \mathbf{Y}_{-n}^1) \right| \\
&\quad = \Big| P_\vartheta(X_k = a \mid X_l = b, \mathbf{Y}_{-n}^1) P_\vartheta(X_l = b \mid \mathbf{Y}_{-n}^1) \\
&\qquad\qquad - P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^1) P_\vartheta(X_l = b \mid \mathbf{Y}_{-n}^1) \Big| \\
&\quad \le \left| P_\vartheta(X_k = a \mid X_l = b, \mathbf{Y}_{-n}^1) - P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^1) \right| \\
&\quad = \Big| \sum_{c=1}^{K} [P_\vartheta(X_k = a \mid X_l = b, \mathbf{Y}_{-n}^1) \\
&\qquad\qquad - P_\vartheta(X_k = a \mid X_l = c, \mathbf{Y}_{-n}^1)] P_\vartheta(X_l = c \mid \mathbf{Y}_{-n}^1) \Big| \\
&\quad \le \max_{a,b,c} \left| P_\vartheta(X_k = a \mid X_l = b, \mathbf{Y}_{-n}^1) - P_\vartheta(X_k = a \mid X_l = c, \mathbf{Y}_{-n}^1) \right| \\
&\quad \le \prod_{i=l+1}^{k-1} \exp(-2\mu_0(Y_i)),
\end{aligned}$$

where the last inequality follows from Lemma 3. The proof with $\mathbf{Y}_{-n}^0$ is analogous. □

Let G denote the neighborhood $\{\vartheta: |\vartheta - \vartheta_0| < \delta\}$ of ϑ_0.

LEMMA 9. *As $m, n \to \infty$,*

$$\left\| \sup_{\vartheta \in G} \left| D^2 \log p_\vartheta(Y_1 | \mathbf{Y}_{-m}^1) - D^2 \log p_\vartheta(Y_1 | \mathbf{Y}_{-n}^1) \right| \right\|_1 \to 0.$$

PROOF. Considering (17), we see that we must prove, for example,

$$\begin{aligned}
\Bigg\| \sup_{\vartheta \in G} \Bigg| \sum_{k=-m}^{0} \sum_{l=-m}^{0} \Big\{ & E_\vartheta[\gamma_\vartheta(Y_k | X_k) \gamma_\vartheta(Y_l | X_l) \mid \mathbf{Y}_{-m}^1] \\
& - E_\vartheta[\gamma_\vartheta(Y_k | X_k) \mid \mathbf{Y}_{-m}^1] E_\vartheta[\gamma_\vartheta(Y_l | X_l) \mid \mathbf{Y}_{-m}^1] \\
& - E_\vartheta[\gamma_\vartheta(Y_k | X_k) \gamma_\vartheta(Y_l | X_l) \mid \mathbf{Y}_{-m}^0] \\
& + E_\vartheta[\gamma_\vartheta(Y_k | X_k) \mid \mathbf{Y}_{-m}^0] E_\vartheta[\gamma_\vartheta(Y_l | X_l) \mid \mathbf{Y}_{-m}^0] \Big\}
\end{aligned} \qquad (18)$$

$$- \sum_{k=-n}^{0} \sum_{l=-n}^{0} \Big\{ E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^{1}] - E_\vartheta[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-n}^{1}] E_\vartheta[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^{1}] - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^{0}] + E_\vartheta[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-n}^{0}] E_\vartheta[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^{0}] \Big\} \Big| \Big\|_1 \to 0$$

as $m, n \to \infty$. Other statements, similar to (18) and which together with (18) prove the lemma, can be shown using slight variations of the technique used below. In order to prove (18), it is sufficient to show that (assuming $m \geq n$) for $j = 0, 1$,

$$\sum_{k=-m}^{-\lfloor n/2 \rfloor} \sum_{l=k}^{\lfloor k/2 \rfloor} \Big\| \sup_{\vartheta \in G} \big| E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-m}^{1}] - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-m}^{0}] \big| \Big\|_1 \to 0, \tag{19}$$

$$\sum_{k=-m}^{-\lfloor n/2 \rfloor} \sum_{l=k}^{\lfloor k/2 \rfloor} \Big\| \sup_{\vartheta \in G} \big| E_\vartheta[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-m}^{1}] E_\vartheta[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-m}^{1}] - E_\vartheta[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-m}^{0}] E_\vartheta[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-m}^{0}] \big| \Big\|_1 \to 0, \tag{20}$$

$$\sum_{k=-n}^{-\lfloor n/2 \rfloor} \sum_{l=k}^{\lfloor k/2 \rfloor} \Big\| \sup_{\vartheta \in G} \big| E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^{1}] - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^{0}] \big| \Big\|_1 \to 0, \tag{21}$$

$$\sum_{k=-n}^{-\lfloor n/2 \rfloor} \sum_{l=k}^{\lfloor k/2 \rfloor} \Big\| \sup_{\vartheta \in G} \big| E_\vartheta[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-n}^{1}] E_\vartheta[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^{1}] - E_\vartheta[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-n}^{0}] E_\vartheta[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^{0}] \big| \Big\|_1 \to 0, \tag{22}$$

$$\sum_{k=-\lfloor n/2 \rfloor}^{0} \sum_{l=-\lfloor n/2 \rfloor}^{0} \Big\| \sup_{\vartheta \in G} \big| E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-m}^{j}] - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^{j}] \big| \Big\|_1 \to 0, \tag{23}$$

$$\sum_{k=-\lfloor n/2 \rfloor}^{0} \sum_{l=-\lfloor n/2 \rfloor}^{0} \Big\| \sup_{\vartheta \in G} \big| E_\vartheta[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-m}^{j}] E_\vartheta[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-m}^{j}] - E_\vartheta[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-n}^{j}] E_\vartheta[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^{j}] \big| \Big\|_1 \to 0, \tag{24}$$

$$\sum_{k=-m}^{-\lfloor n/2\rfloor}\sum_{l=-\lfloor k/2\rfloor}^{0}\Big\|\sup_{\vartheta\in G}\big|E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-m}^{j}] - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-m}^{j}]E_\vartheta[\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-m}^{j}]\big|\Big\|_1\to 0, \tag{25}$$

$$\sum_{k=-n}^{-\lfloor n/2\rfloor}\sum_{l=-\lfloor k/2\rfloor}^{0}\Big\|\sup_{\vartheta\in G}\big|E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-n}^{j}] - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^{j}]E_\vartheta[\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-n}^{j}]\big|\Big\|_1\to 0, \tag{26}$$

as $m, n\to\infty$; compare Figure 1. The idea of splitting up the sum (18) goes back to Baum and Petrie (1966).

Starting with (19), by the first part of Lemma 7 we have that

$$\begin{aligned}
&\sup_{\vartheta\in G}\big|E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-m}^{1}] - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-m}^{0}]\big|\\
&\quad\le \sup_{\vartheta\in G}\sum_{a,b=1}^{K}|\gamma_\vartheta(Y_k|a)|\,|\gamma_\vartheta(Y_l|b)|\,\big|P_\vartheta(X_k=a,\ X_l=b\mid \mathbf{Y}_{-m}^{1})\\
&\qquad\qquad - P_\vartheta(X_k=a, X_l=b\mid \mathbf{Y}_{-m}^{0})\big|\\
&\quad\le C\Big(\sup_{\vartheta\in G}\max_a|\gamma_\vartheta(Y_k|a)|\Big)\Big(\sup_{\vartheta\in G}\max_a|\gamma_\vartheta(Y_l|b)|\Big)\prod_{i=k\vee l+1}^{0}\exp(-2\mu_0(Y_i)).
\end{aligned}$$

By conditioning on the X's, we obtain that the $\mathbb{L}_1(P_0)$-norm of the above expression is bounded by $C\beta^{|k|\wedge|l|}$ for some $\beta\in[0,1)$, whence the left-hand

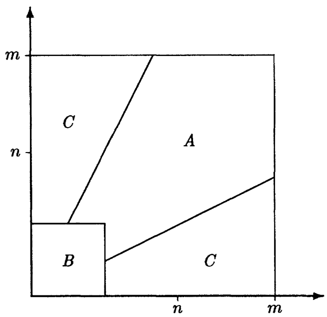


FIG. 1. *Illustration of how the sum in* (18) *is split into subregions. In region A,* $E_\vartheta[\cdot\mid \mathbf{Y}_{-m}^{1}]$ *is compared to* $E_\vartheta[\cdot\mid \mathbf{Y}_{-m}^{0}]$ *etc. In region B,* $E_\vartheta[\cdot\mid \mathbf{Y}_{-m}^{1}]$ *is compared to* $E_\vartheta[\cdot\mid \mathbf{Y}_{-n}^{1}]$ *etc. In region C,* $E_\vartheta[\cdot\times\cdot\mid \mathbf{Y}_{-m}^{1}]$ *is compared to* $E_\vartheta[\cdot\mid \mathbf{Y}_{-m}^{1}]\times E_\vartheta[\cdot\mid \mathbf{Y}_{-m}^{1}]$ *and so on.*

side of (19) is bounded by

$$C \sum_{k=\lfloor n/2 \rfloor}^{m} \sum_{l=\lfloor k/2 \rfloor}^{m} \beta^l \le C \sum_{k=\lfloor n/2 \rfloor}^{m} \beta^{\lfloor k/2 \rfloor} \le C\beta^{\lfloor n/4 \rfloor}.$$

Here, the right-hand side tends to zero as $m, n \to \infty$, and (19) follows; (21) follows similarly.

For (20), the first part of Lemma 5 shows that for any $\vartheta \in G$,

$$\begin{aligned}
&\max_{a,b}\big|P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^1)P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^1) \\
&\qquad - P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^0)P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^0)\big| \\
&\quad \le \max_{a,b}\big|P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^1)P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^1) \\
&\qquad\qquad - P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^1)P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^0) \\
&\qquad\qquad + P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^1)P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^0) \\
&\qquad\qquad - P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^0)P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^0)\big| \\
&\quad \le \max_{b}|P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^1) - P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^0)| \\
&\qquad + \max_{a}|P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^1) - P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^0)| \\
&\quad \le 2 \prod_{i=k\vee l+1}^{0} \exp(-2\mu_0(Y_i)),
\end{aligned} \tag{27}$$

so that

$$\begin{aligned}
&\sup_{\vartheta\in G}\big|E_\vartheta[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-m}^1]E_\vartheta[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-m}^1] \\
&\qquad - E_\vartheta[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-m}^0]E_\vartheta[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-m}^0]\big| \\
&\quad \le \sup_{\vartheta\in G} \sum_{a,b=1}^{K} |\gamma_\vartheta(Y_k|a)|\,|\gamma_\vartheta(Y_l|b)| \\
&\qquad\qquad \times \big|P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^1)P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^1) \\
&\qquad\qquad\quad - P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^0)P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^0)\big| \\
&\quad \le C\Big(\sup_{\vartheta\in G}\max_a |\gamma_\vartheta(Y_k|a)|\Big)\Big(\sup_{\vartheta\in G}\max_a |\gamma_\vartheta(Y_l|a)|\Big) \prod_{i=k\vee l+1}^{0} \exp(-2\mu_0(Y_i)).
\end{aligned}$$

Now (20) follows as above, and (22) follows similarly.

Further, the second part of Lemma 7 shows that the left-hand side of (23) is bounded by

$$\begin{aligned} C \sum_{k=-\lfloor n/2 \rfloor}^{0} \sum_{l=-\lfloor n/2 \rfloor}^{0} \beta^{n+k\wedge l-1} &= C \sum_{k=0}^{\lfloor n/2 \rfloor} \sum_{l=0}^{\lfloor n/2 \rfloor} \beta^{n-k\vee l-1} \\ &\le 2C \sum_{k=0}^{\lfloor n/2 \rfloor} \sum_{l=k}^{\lfloor n/2 \rfloor} \beta^{n-l-1} \\ &\le C \sum_{k=0}^{\lfloor n/2 \rfloor} \beta^{\lfloor n/2 \rfloor} \le C(\lfloor n/2 \rfloor + 1)\beta^{\lfloor n/2 \rfloor}. \end{aligned}$$

The right-hand side vanishes as $n \to \infty$, whence (23) follows; (24) follows using a bound similar to (27).

Finally, by Lemma 8 the left-hand side of (25) is bounded by

$$\begin{aligned} C \sum_{k=-m}^{-\lfloor n/2 \rfloor} \sum_{l=\lfloor k/2 \rfloor}^{0} \beta^{k\vee l-k\wedge l-1} &= C \sum_{k=\lfloor n/2 \rfloor}^{m} \sum_{l=0}^{\lfloor k/2 \rfloor} \beta^{k\vee l-k\wedge l-1} \\ &= C \sum_{k=\lfloor n/2 \rfloor}^{m} \sum_{l=0}^{\lfloor k/2 \rfloor} \beta^{k-l-1} \\ &\le C \sum_{k=\lfloor n/2 \rfloor}^{m} \beta^{k-\lfloor k/2 \rfloor-1} \le C\beta^{\lfloor n/4 \rfloor}, \end{aligned}$$

whence (25) follows; (26) follows similarly, and the proof is complete. $\square$

Thus, $\{D^2 \log p_\vartheta(Y_1|Y_0, \ldots, Y_{-n})\}$ is a "uniform Cauchy sequence" in $\mathbb{L}_1(P_0)$, and the following result is then immediate.

LEMMA 10. *There is a continuous function $\zeta_1(\vartheta)$ from G to $\mathbb{L}_1(P_0)$ such that*

$$\left\| \sup_{\vartheta \in G} |D^2 \log p_\vartheta(Y_1|Y_0, \ldots, Y_{-n}) - \zeta_1(\vartheta)| \right\|_1 \to 0$$

as $n \to \infty$.

REMARK. Assuming the MLE to be consistent, that is, that (A6) holds, any subset of the sample space with P_ϑ-measure one for some $\vartheta \neq \vartheta_0$ has P_0-measure zero, whence Lemma 5 does not guarantee that any of the statements with infinite n or m holds P_0-a.s. for any ϑ other than ϑ_0. This is the reason for working with Cauchy sequences in this section, rather than with an explicit representation of $\zeta_1(\vartheta)$ similar to (6).

PROOF OF LEMMA 2. Define $\zeta_k(\vartheta)$ as the $\mathbb{L}_1(P_0)$-limit of

$$D^2 \log p_\vartheta(Y_k|\mathbf{Y}_{-n}^{k-1})$$

and let G' be an arbitrary neighborhood of ϑ_0 such that $G' \subseteq G$. We then have

$$
\begin{aligned}
&\limsup_{n\to\infty} P_0\left(\left|n^{-1}\ddot{L}_n(\vartheta_n^*) - n^{-1}\sum_{k=1}^n \zeta_k(\vartheta_0)\right| > \varepsilon\right)\\
&\quad = \limsup_{n\to\infty} P_0\left(\left|n^{-1}\sum_{k=1}^n \{D^2 \log p_{\vartheta_n^*}(Y_k|Y_{k-1},\dots,Y_1) - \zeta_k(\vartheta_0)\}\right| > \varepsilon\right)\\
&\quad \le \limsup_{n\to\infty} P_0\left(n^{-1}\sum_{k=1}^n \sup_{\vartheta\in G'}\left|D^2 \log p_{\vartheta}(Y_k|Y_{k-1},\dots,Y_1) - \zeta_k(\vartheta_0)\right| > \varepsilon\right)\\
&\qquad + \limsup_{n\to\infty} P_0(\vartheta_n^* \notin G')\\
&\quad \le \limsup_{n\to\infty} n^{-1}\varepsilon^{-1}\sum_{k=1}^n \left\|\sup_{\vartheta\in G'}\left|D^2 \log p_{\vartheta}(Y_1|Y_0,\dots,Y_{-k+2}) - \zeta_1(\vartheta_0)\right|\right\|_1\\
&\quad \le \limsup_{n\to\infty} n^{-1}\varepsilon^{-1}\sum_{k=1}^n \left\|\sup_{\vartheta\in G'}\left|D^2 \log p_{\vartheta}(Y_1|Y_0,\dots,Y_{-k+2}) - \zeta_1(\vartheta)\right|\right\|_1\\
&\qquad + \limsup_{n\to\infty} n^{-1}\varepsilon^{-1}\sum_{k=1}^n \left\|\sup_{\vartheta\in G'}\left|\zeta_1(\vartheta) - \zeta_1(\vartheta_0)\right|\right\|_1\\
&\quad = \varepsilon^{-1}\left\|\sup_{\vartheta\in G'}\left|\zeta_1(\vartheta) - \zeta_1(\vartheta_0)\right|\right\|_1,
\end{aligned}
$$

where the third step follows by Markov's inequality and stationarity, and the last one by Lemma 10. Let $G' \downarrow \{\vartheta_0\}$ and use continuity of $\zeta(\cdot)$ to conclude that

$$n^{-1}\ddot{L}_n(\vartheta^*) - n^{-1}\sum_{k=1}^n \zeta_k(\vartheta_0) \to 0 \text{ in } P_0\text{-probability} \tag{28}$$

as $n \to \infty$.

Now, because $\{Y_k\}$ is ergodic, so is $\{\zeta_k(\vartheta_0)\}$, whence $n^{-1}\sum_1^n \zeta_k(\vartheta_0) \to J$ P_0-a.s. for some matrix $J = E_0\zeta_1(\vartheta_0)$. The proof is thus complete if we can show that $J = -\mathscr{I}_0$.

Using (A3)(iii) it readily follows that

$$E_0[-D^2 \log g_{\vartheta_0}(Y_1|X_1)] = E_0[(D\log g_{\vartheta_0}(Y_1|X_1))^2],$$

which together with the representations (4) and (17) show that

$$E_0[D^2 \log p_{\vartheta_0}(Y_1|Y_0,\dots,Y_{-n})] = -E_0\left[\left(D\log p_{\vartheta_0}(Y_1|Y_0,\dots,Y_{-n})\right)^2\right]$$

for each n. Hence, by Lemma 6 and Lemma 10, $J = -\mathscr{I}_0$. □

Acknowledgments. Many thanks to Jens Ledet Jensen and Niels Væver Petersen, who did not only carefully read an earlier version of this paper and found four errors (in Assumption A2 and the proofs of Lemmas 1, 2 and 9), but who also provided solutions to these errors.

REFERENCES

ALBERT, P. S. (1991). A two-state Markov mixture model for a time series of epileptic seizure counts. *Biometrics* **47** 1371–1381.

BAUM, L. E. and PETRIE, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.* **37** 1554–1563.

BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41** 164–171.

BICKEL, P. J. and RITOV, Y. (1996). Inference in hidden Markov models I: local asymptotic normality in the stationary case. *Bernoulli* **2** 199–228.

DURRETT, R. (1991). *Probability: Theory and Examples*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

FREDKIN, D. R. and RICE, J. A. (1992). Maximum likelihood estimation and identification directly from single-channel recordings. *Proc. Royal Soc. London Ser. B* **249** 125–132.

GUTTORP, P. (1995). *Stochastic Modeling of Scientific Data*. Chapman & Hall, London.

HEFFES, H. and LUCANTONI, D. (1986). A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE J. Select. Areas Comm.* **4** 856–867.

JAMSHIDIAN, M. and JENNRICH, R. I. (1997). Acceleration of the EM algorithm by using quasi-Newton methods. *J. Royal Statist. Soc. Ser. B* **59** 569–587.

LE, N. D., LEROUX, B. G. and PUTERMAN, M. L. (1992). Reader reaction: exact likelihood evaluation in a Markov mixture model for time series of seizure counts. *Biometrics* **48** 317–323.

LEROUX, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic Process. Appl.* **40** 127–143.

LEROUX, B. G. and PUTERMAN, M. L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* **48** 545–558.

LINDGREN, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scand. J. Statist.* **5** 81–91.

LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Royal Statist. Soc. Ser. B* **44** 226–233.

MACDONALD, I. L. and ZUCCHINI, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman & Hall, London.

MCLACHLAN, G. J. and KRISHNAN, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.

MENG, X.-L. and VAN DYK, D. (1997). The EM algorithm—an old folk-song sung to a new fast tune (with discussion). *J. Royal Statist. Soc. Ser. B* **59** 511–567.

PETRIE, T. (1969). Probabilistic functions of finite state Markov chains. *Ann. Math. Statist.* **40** 97–115.

PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. and VETTERLING, W. T. (1989). *Numerical Recipes*. Cambridge Univ. Press.

RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77** 257–284.

RITOV, Y. (1996). Uniform convergence of quasi-convex functions with applications to missing data and hidden Markov models. Preprint.

RYDÉN, T. (1994). Parameter estimation for Markov modulated Poisson processes. *Stochastic Models* **10** 795–829.

SHIRYAYEV, A. N. (1984). *Probability*. Springer, New York.

P. J. BICKEL
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
EVANS HALL
BERKELEY, CALIFORNIA 94720

Y. RITOV
DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY
JERUSALEM 91905
ISRAEL

T. RYDÉN
DEPARTMENT OF MATHEMATICAL STATISTICS
LUND UNIVERSITY
BOX 118
S-221 00 LUND
SWEDEN
E-MAIL: tobias@maths.lth.se

Chapter 4
Function Estimation

Hans-Georg Müller

4.1 Introduction to Three Papers on Nonparametric Curve Estimation

4.1.1 Introduction

The following is a brief review of three landmark papers of Peter Bickel on theoretical and methodological aspects of nonparametric density and regression estimation and the related topic of goodness-of-fit testing, including a class of semiparametric goodness-of-fit tests. We consider the context of these papers, their contribution and their impact. Bickel's first work on density estimation was carried out when this area was still in its infancy and proved to be highly influential for the subsequent wide-spread development of density and curve estimation and goodness-of-fit testing.

The first of Peter Bickel's contributions to kernel density estimation was published in 1973, nearly 40 years ago, when the field of nonparametric curve estimation was still in its infancy and was poised for the subsequent rapid expansion, which occurred later in the 1970s and 1980s. Bickel's work opened fundamental new perspectives, that were not fully developed until much later. Kernel density estimation was formalized in Rosenblatt (1956) and then developed further in Parzen (1962), where bias expansions and other basic techniques for the analysis of these nonparametric estimators were showcased.

Expanding upon an older literature on spectral density estimation, this work set the stage for substantial developments in nonparametric curve estimation that began in the later 1960s. This earlier literature on curve estimation is nicely surveyed in Rosenblatt (1971) and it defined the state of the field when Peter Bickel made

H.-G. Müller (✉)
Department of Statistics, University of California, Davis, CA, USA
e-mail: hgmueller@ucdavis.edu

J. Fan et al. (eds.), *Selected Works of Peter J. Bickel*, Selected Works in Probability and Statistics 13, DOI 10.1007/978-1-4614-5544-8_4,

the first highly influential contribution to nonparametric curve estimation in Bickel and Rosenblatt (1973). This work not only connected for the first time kernel density estimation with goodness-of-fit testing, but also did so in a mathematically elegant way.

A deep study of the connection between smoothness and rates of convergence and improved estimators of functionals of densities, corresponding to integrals of squared derivatives, is the hallmark of Bickel and Ritov (1988). Estimation of these density functionals has applications in determining the asymptotic variance of nonparametric location statistics. Functional of this type also appear as a factor in the asymptotic leading bias squared term for the mean integrated squared error. Thus the estimation of these functional has applications for the important problem of bandwidth choice for nonparametric kernel density estimates.

In the third article covered in this brief review, Bickel and Li (2007) introduce a new perspective to the well-known curse of dimensionality that affects any form of smoothing and nonparametric function estimation in high dimension: It is shown that for local linear smoothers in a nonparametric regression setting where the predictors at least locally lie on an unknown manifold, the curse of dimensionality effectively is not driven by the ostensible dimensionality of the predictors but rather by the dimensionality of the predictors, which might be much lower. In the case of relatively low-dimensional underlying manifolds, the good news is that the curse would then not be as severe as it initially appears, and one may obtain unexpectedly fast rates of convergence.

The first two papers that are briefly discussed here create a bridge between density estimation and goodness-of-fit. The goodness-of-fit aspect is central to Bickel and Rosenblatt (1973), while a fundamental transition phenomenon and improved estimation of density functionals are key aspects of Bickel and Ritov (1988). Both papers had a major impact in the field of nonparametric curve estimation. The third paper (Bickel and Li 2007) creates a fresh outlook on nonparametric regression and will continue to inspire new approaches. Some remarks on Bickel and Rosenblatt (1973) can be found in Sect. 2, on Bickel and Ritov (1988) Sect. 3, and on Bickel and Li (2007) in Sect. 4.

4.1.2 Density Estimation and Goodness-of-Fit

Nonparametric curve estimation originated in spectral density estimation, where it had been long known that smoothing was mandatory to improve the properties of such estimates (Daniell 1946; Einstein 1914). The smoothing field expanded to become a major field in nonparametric statistics around the time the paper Bickel and Rosenblatt (1973) appeared. At that time, kernel density estimation and other basic nonparametric estimators of density functions such as orthogonal least squares (Čencov 1962) were established. While many results were available in 1973 about local properties of these estimates, there had been no in-depth investigation yet of their global behavior.

This is where Bickel's influential contribution came in. Starting with the Rosenblatt-Parzen kernel density estimator

$$f_n(x) = \frac{1}{nb(n)} \sum_{i=1}^{n} w\left(\frac{x - X_i}{b(n)}\right) = \int \frac{1}{b(n)} w\left(\frac{x-u}{b(n)}\right) dF_n(u), \tag{4.1}$$

where $b(n)$ is a sequence of bandwidths that converges to 0, but not too fast, w a kernel function and dF_n stands for the empirical measure, Bickel and Rosenblatt (1973) consider the functionals

$$D_1 = \sup_{a_1 \le x \le a_2} |f_n(x) - f(x)| / (f(x))^{1/2}, \tag{4.2}$$

$$D_2 = \int_{a_1}^{a_2} \frac{[f_n(x) - f(x)]^2}{f(x)}. \tag{4.3}$$

The asymptotic behavior of these two functionals proves to be quite different. Functional D_1 corresponds to a maximal deviation on the interval, while functional D_2 is an integral and can be interpreted as a weighted integrated absolute deviation. While D_2, properly scaled, converges to a Gaussian limit, D_1 converges to an extreme value distribution. Harnessing the maximal deviation embodied in D_1 was the first serious attempt to obtain global inference in nonparametric density estimation. As Bickel and Rosenblatt (1973) state, *the statistical interest in this functional is twofold, as (i) a convenient way of getting a confidence band for f. (ii) A test statistic for the hypothesis $H_0 : f = f_0$.* They thereby introduce the goodness-of-fit theme, that constitutes one major motivation for density estimation and has spawned much research to this day. Motivation (i) leads to Theorem 3.1, and (ii) to Theorem 3.2 in Bickel and Rosenblatt (1973).

In their proofs, Bickel and Rosenblatt (1973) use a strong embedding technique, which was quite recent at the time. Theorem 3.1 is a remarkable achievement. If one employs a rectangular kernel function $w = 1_{[-\frac{1}{2},\frac{1}{2}]}$ and a bandwidth sequence $b(n) = n^{-\delta}$, $0 < \delta < \frac{1}{2}$, then the result in Theorem 3.1 is for centered processes

$$P\left[(2\delta \log n)^{1/2}\left([nb(n)f^{-1}(t)]^{1/2} \sup_{a_1,a_2}[f_n(t) - E(f_n(t))] - d_n\right) < x\right] \to e^{-2e^{-x}},$$

where

$$d_n = \rho_n - \frac{1}{2}\rho_n^{-1}[\log(\pi + \delta) + \log\log n], \quad \rho_n = (2\delta \log n)^{1/2}.$$

The slow convergence to the limit that is indicated by the rate $(\log n)^{1/2}$ is typical for maximal deviation results in curve estimation, of which Theorem 3.1 is the first. A multivariate version of this result appeared in Rosenblatt (1976).

A practical problem that has been discussed by many authors in the 1980s and 1990s has been how to handle the bias for the construction of confidence intervals and density-estimation based inference in general. This is a difficult problem. It is also related to the question how one should choose bandwidths when constructing confidence intervals, even pointwise rather than global ones, in relation to choosing the bandwidth for the original curve estimate for which the confidence region is desired (Hall 1992; Müller et al. 1987). For instance, undersmoothing has been advocated and also other specifically designed bias corrections. This is of special relevance when the maximal deviation is to be constructed over intervals that include endpoints of the density, where bias is a particularly notorious problem.

For inference and goodness-of-fit testing, Bickel and Rosenblatt (1973), based on the deviation D_2 as in (4.3), propose the test statistic

$$T_n = \int [f_n(x) - E(f_n(x))]^2 a(x)\, dx$$

with a weight function a for testing the hypothesis H_0. Compared to classical goodness-of-fit tests, this test is shown to be better than the χ^2 test and incorporates nuisance parameters as needed. This Bickel-Rosenblatt test has encountered much interest; an example is an application for testing independence (Rosenblatt 1975).

Recent extensions and results under weaker conditions include extensions to the case of an error density for stationary linear autoregressive processes that were developed in Lee and Na (2002) and Bachmann and Dette (2005), and for GARCH processes in Koul and Mimoto (2010). A related L^1-distance based goodnes-of-fit test was proposed in Cao and Lugosi (2005), while a very general class of semiparametric tests targeting composite hypotheses was introduced in Bickel et al. (2006).

4.1.3 Estimating Functionals of a Density

Kernel density estimators (4.1) require specification of a kernel function w and of a bandwidth or smoothing parameter $b = b(n)$. If one uses a kernel function that is a symmetric density, this selection can be made based on the asymptotically leading term of mean integrated squared error (MISE),

$$\frac{1}{4} b(n)^4 \int w(u) u^2\, du \int [f^{(2)}(x)]^2\, dx + [nb(n)]^{-1} \int w(u)^2\, du,$$

which leads to the asymptotically optimal bandwidth

$$b^*(n) = c \left(n \int [f^{(2)}(x)]^2\, dx \right)^{-1/5},$$

where c is a known constant. In order to determine this optimal bandwidth, one is therefore confronted with the problem of estimating integrated squared density derivatives

$$\int [f^{(k)}(x)]^2 \, dx, \tag{4.4}$$

where cases $k > 2$ are of interest when choosing bandwidths for density estimates with higher order kernels. These have faster converging bias at the cost of increasing variance but are well known to have rates of convergence that are faster in terms of MISE, if the underlying density is sufficiently smooth and optimal bandwidths are used. Moreover, the case $k = 0$ plays a role in the asymptotic variance of rank-based estimators (Schweder 1975).

The relevance of the problem of estimating density functionals of type (4.4) had been recognized by various authors, including Hall and Marron (1987), at the time the work Bickel and Ritov (1988) was published. The results of Bickel and Ritov however are not a direct continuation of the previous line of research; rather, they constitute a surprising turn of affairs. First, the problem is positioned within a more general semiparametric framework. Second, it is established that the $\sqrt{n}$ of convergence that one expects for functionals of type (4.4) holds if $f^{(m)}$ is Hölder continuous of order α with $m + \alpha > 2k + \frac{1}{4}$, and, with an element of surprise, that it does not hold in a fairly strong sense when this condition is violated.

The upper bound for this result is demonstrated by utilizing kernel density estimates (4.1), employing a kernel function of order $\max(k, m - k) + 1$ and then using plug-in estimators. However, straightforward plug-in estimators suffer from bias that is severe enough to prevent optimal results. Instead, Bickel and Ritov employ a clever bias correction term (that appears in their equation (2.2) after the plug-in estimator is introduced) and then proceed to split the sample into two separate parts, combining two resulting estimators.

An amazing part of the paper is the proof that an unexpected and surprising phase transition occurs at $\alpha = 1/4$. This early example for such a phase transition hinges on an ingenious construction of a sequence of measures and the Bayes risk for estimating the functional. For less smooth densities, where the transition point has not been reached, Bickel and Rosenblatt (1973) provide the optimal rate of convergence, a rate slower than $\sqrt{n}$. The arguments are connected more generally with semiparametric information bounds in the precursor paper Bickel (1982).

Bickel and Ritov (1988) is a landmark paper on estimating density functionals that inspired various subsequent works by other authors. These include further study of aspects that had been left open, such as adaptivity of the estimators (Efromovich and Low 1996), extensions to more general density functionals with broad applications (Birgé and Massart 1995) and the study of similar problems for other curve functionals, for example integrated second derivative estimation in nonparametric regression (Efromovich and Samarov 2000).

4.1.4 Curse of Dimensionality for Nonparametric Regression on Manifolds

It has been well known since Stone (1980) that all nonparametric curve estimation methods, including nonparametric regression and density estimation, suffer severely in terms of rates of convergence in high-dimensional or even moderately dimensioned situations. This is born out in statistical practice, where unrestricted nonparametric curve estimation is known to make little sense if moderately sized data have predictors with dimensions say $D \geq 4$. Assuming the function to be estimated is in a Sobolev space of smoothness p, optimal rates of convergence of Mean Squared Error and similar measures are $n^{-2p/(2p+D)}$ for samples of size n. To circumvent the curse of dimensionality, alternatives to unrestricted nonparametric regression have been developed, ranging from additive, to single index, to additive partial linear models. Due to their inherent structural constraints, such approaches come at the cost of reduced flexibility with the associated risk of increased bias.

The cause of the curse of dimensionality is the trade-off between bias and variance in nonparametric curve estimation. Bias control demands to consider data in a small neighbourhood around the target predictor levels $\mathbf{x}$, where the curve estimate is desired, while variance control requires large neighbourhoods containing many predictor-response pairs. For increasing dimensions, the predictor locations become increasingly sparse, with larger average distances between predictor locations, moving the variance-bias trade-off and resulting rate of convergence in an unfavorable direction.

Using an example where $p = 2$ and the local linear regression method, Bickel and Li (2007) analyze what happens if the predictors are in fact not only located on a compact subset of $\mathscr{R}^D$, where D is potentially large, but in fact are, at least locally around $\mathbf{x}$, located on a lower-dimensional manifold with intrinsic dimension $d < D$. They derive that in this situation, one obtains the better rate $n^{-2p/(2p+d)}$, where the manifold is assumed to satisfy some local regularity conditions, but otherwise is unknown. This can lead to dramatic gains in rates of convergence, especially if $d = 1, 2$ while D is large.

This nice result can be interpreted as a consequence of the denser packing of the predictors on the lower-dimensional manifold with smaller average distances as compared to the average distances one would expect for the ostensible dimension D of the space, when the respective densities are not degenerate. A key feature is that knowledge of the manifold is not needed to take advantage of its presence. The data do not even have to be located precisely on the manifold, as long as their deviation from the manifold becomes small asymptotically. Bickel and Li (2007) also provide thoughtful approaches to bandwidth choices for this situation and for determining the intrinsic dimension of the unknown manifold, and thus the rate of effective convergence that is determined by d.

This approach likely will play an important role in the ongoing intensive quest for flexible yet fast converging dimension reduction and regression models. Methods for variable selection, dimension reduction and for handling collinearity among predictors, as well as extensions to "large p, small n" situations are in high demand. The idea of exploiting underlying manifold structure in the predictor space for these purposes is powerful, as has been recently demonstrated in Mukherjee et al. (2010) and Aswani et al. (2011). These promising approaches define a new line of research for high-dimensional regression modeling.

Acknowledgements Supported in part by NSF grant DMS-1104426.

References

Aswani A, Bickel P, Tomlin C (2011) Regression on manifolds: estimation of the exterior derivative. Ann Stat 39:48–81

Bachmann, D, Dette H (2005) A note on the Bickel-Rosenblatt test in autoregressive time series. Stat Probab Lett 74:221–234

Bickel P (1982) On adaptive estimation. Ann Stat 10:647–671

Bickel P, Li B (2007). Local polynomial regression on unknown manifolds. In: Complex datasets and inverse problems: tomography, networks and beyond. IMS lecture notes-monograph series, vol 54. Institute of Mathematical Statistics, Beachwood, pp 177–186

Bickel P, Ritov Y (1988) Estimating integrated squared densiuty derivatives: sharp best order of convergence estimates. Sankhya Indian J Stat Ser A 50:381–393

Bickel P, Rosenblatt M (1973) On some global measures of the deviations of density function estimates. Ann Stat 1:1071–1095

Bickel P, Ritov Y, Stoker T (2006) Tailor-made tests for goodness of fit to semiparametric hypotheses. Ann Stat 34:721–741

Birgé L, Massart P (1995) Estimation of integral functionals of a density. Ann Stat 23:11–29

Cao R, Lugosi G (2005) Goodness-of-fit tests based on kernel density estimator. Scand J Stat 32:599–616

Čencov N (1962) Evaluation of an unknown density from observations. Sov Math 3:1559–1562

Daniell P (1946) Discussion of paper by M.S. Bartlett. J R Stat Soc Suppl 8:88–90

Efromovich S, Low M (1996) On Bickel and Ritov's conjecture about adaptive estimation of the integral of the square of density derivative. Ann Stat 24:682–686

Efromovich S, Samarov A (2000) Adaptive estimation of the integral of squared regression derivatives. Scand J Stat 27:335–351

Einstein A (1914) Méthode pour la détermination de valeurs statistiques d'observations concernant des grandeurs soumises à des fluctuations irrégulières. Arch Sci Phys et Nat Ser 4 37:254–256

Hall P (1992) Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. Ann Stat 20:675–694

Hall P, Marron J (1987) Estimation of integrated squared density derivatives. Stat Probab Lett 6:109–115

Koul H, Mimoto N (2010) A goodness-of-fit test for garch innovation density. Metrika 71:127–149

Lee S, Na S (2002) On the Bickel-Rosenblatt test for first order autoregressive models. Stat Probab Lett 56:23–35

Mukherjee S, Wu Q, Zhou D (2010) Learning gradients on manifolds. Bernoulli 16:181–207

Müller H-G, Stadtmüller U, Schmitt T (1987) Bandwidth choice and confidence intervals for derivatives of noisy data. Biometrika 74:743–749

Parzen E (1962) On estimation of a probability density function and mode. Ann Math Stat 33:1065–1076

Rosenblatt M (1956) Remarks on some nonparametric estimates of a density function. Ann Math Stat 27:832–837

Rosenblatt M (1971) Curve estimates. Ann Stat 42:1815–1842
Rosenblatt M (1975) A quadratic measure of deviation of two-dimensional density estimates and a test of independence. Ann Stat 3:1–14
Rosenblatt M (1976) On the maximal deviation of k-dimensional density estimates. Ann Probab 4:1009–1015
Schweder T (1975) Window estimation of the asymptotic variance of rank estimators of location. Scand J Stat 2:113–126
Stone CJ (1980) Optimal rates of convergence for nonparametric estimators. Ann Stat 10:1040–1053

The Annals of Statistics
1973, Vol. 1, No. 6, 1071-1095

ON SOME GLOBAL MEASURES OF THE DEVIATIONS OF DENSITY FUNCTION ESTIMATES

BY P. J. BICKEL[1] AND M. ROSENBLATT[2]

University of California, Berkeley; University of California, San Diego

We consider density estimates of the usual type generated by a weight function. Limt theorems are obtained for the maximum of the normalized deviation of the estimate from its expected value, and for quadratic norms of the same quantity. Using these results we study the behavior of tests of goodness-of-fit and confidence regions based on these statistics. In particular, we obtain a procedure which uniformly improves the chi-square goodness-of-fit test when the number of observations and cells is large and yet remains insensitive to the estimation of nuisance parameters. A new limit theorem for the maximum absolute value of a type of nonstationary Gaussian process is also proved.

1. Introduction. Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed random variables with continuous density function $f(x)$. By now there are a goodly number of papers on estimation of the density function (see [13] for a bibliography). The class of estimates $f_n(x)$ that we consider are determined by a bounded integrable weight function w,

$$f_n(x) = \frac{1}{nb(n)} \sum_{j=1}^n w\left(\frac{x - X_j}{b(n)}\right) \tag{1.1}$$
$$= \int \frac{1}{b(n)} w\left(\frac{x-s}{b(n)}\right) dF_n(s) .$$

In formula (1.1), F_n is the sample distribution function. Also $b(n)$ is a bandwidth that tends to zero as $n \to \infty$ but is such that $n^{-1} = o(b(n))$.

The local properties of such estimates have been discussed extensively. Our object will be to get global measures of how good $f_n(x)$ is as an estimate of $f(x)$. In particular, the asymptotic distribution of the functionals

$$\max_{0 \leq x \leq 1} |f_n(x) - f(x)|/(f(x))^{\frac{1}{2}}$$

and

$$\int_0^1 \frac{[f_n(x) - f(x)]^2}{f(x)} dx$$

are evaluated under appropriate conditions as $n \to \infty$.

Received November 1971; revised March 1973.

[1] This work was performed while the author was a fellow of the John Simon Guggenheim Memorial Foundation and supported in part by the Office of Naval Research under Contract N00014-69-A0200-1038.

[2] This work was carried out while the author was a fellow of the John Simon Guggenheim Memorial Foundation and supported in part by the Office of Naval Research.

Key words and phrases. Density function estimates, weight function, bandwidth, global measure of deviation, asymptotic distribution, functionals of stochastic process, Gaussian process.

We shall state two results, one concerned with absolute deviation of the estimate $f_n(x)$ from $f(x)$ and the other with integrated quadratic deviation. They will give some insight into the type of result that is obtained. However, in order to give the result on absolute deviation it is convenient to introduce at this point certain convenient assumptions which we shall refer to as A1, A2, A3, and A4.

A1. The weight function w also assigns mass one to the line and either (a) vanishes outside an interval $[-A, A]$ and is absolutely continuous on $[-A, A]$ with derivative w' or (b) is absolutely continuous on $(-\infty, \infty)$ with derivative w' such that $\int |w'(t)|^k \, dt < \infty$, $k = 1, 2$.

A2. The density f is continuous, positive and bounded.

A3. The function $f^{\frac{1}{2}}$ is absolutely continuous and its derivative $\frac{1}{2}f'/f^{\frac{1}{2}}$ is bounded in absolute value. Moreover,

$$\int_{[|z| \geq 3]} |z|^{\frac{3}{2}} [\log \log |z|]^{\frac{1}{2}} [|w'(z)| + |w(z)|] \, dz < \infty .$$

A4. The second derivative f'' of f exists and is bounded. Moreover w is symmetric (about 0) and $z^2 w(z)$ is integrable.

We shall simply state a corollary of a main result on absolute deviations which is appealing because it is phrased in a form that is convenient if one wishes to set up a confidence band for the density function.

COROLLARY. *Let assumptions* A1—A4 *be satisfied with* $b(n) = n^{-\delta}$, $\frac{1}{5} < \delta < \frac{1}{2}$. *Then*

$$\begin{aligned} \lim_{n \to \infty} P\Bigg[f_n(x) - \left(\frac{f_n(x)\lambda(w)}{nb(n)} \right)^{\frac{1}{2}} \left(\frac{z}{(2\delta \log n)^{\frac{1}{2}}} + d_n \right) & \\ \leq f(x) \leq f_n(x) + \left(\frac{f_n(x)\lambda(w)}{nb(n)} \right)^{\frac{1}{2}} \left(\frac{z}{(2\delta \log n)^{\frac{1}{2}}} + d_n \right) \ \textit{for all} \ 0 \leq x \leq 1 \Bigg] & \\ = e^{-2e^{-z}} & \end{aligned} \tag{1.2}$$

where

$$\lambda(w) = \int w^2(t) \, dt$$

and

$$d_n = (2\delta \log n)^{\frac{1}{2}} + \frac{1}{(2\delta \log n)^{\frac{1}{2}}} \left\{ \log \left(\frac{K_1(w)}{\pi^{\frac{1}{2}}} \right) + \tfrac{1}{2}[\log \delta + \log \log n] \right\}$$

if (a) *of* A1 *holds and*

$$K_1(w) = \frac{w^2(A) + w^2(-A)}{2} \Big/ \lambda(w) > 0 ,$$

and otherwise

$$d_n = (2\delta \log n)^{\frac{1}{2}} + \frac{1}{(2\delta \log n)^{\frac{1}{2}}} \left[\log \frac{1}{\pi} \left(\frac{K_2(w)}{2} \right)^{\frac{1}{2}} \right]$$

with

$$K_2(w) = \tfrac{1}{2} \int_{-\infty}^{\infty} [w'(t)]^2 \, dt / \lambda(w) .$$

The following result for a quadratic functional is also of some interest. The function $a(x)$ used in the theorem is assumed to be a bounded piece-wise smooth integrable function.

THEOREM. *Let* A1—A4 *hold. Then if* $b(n) = o(n^{-\frac{2}{9}})$, $n^{-\frac{1}{4}}(\log n)^{\frac{1}{2}}(\log\log n)^{\frac{1}{4}} = o(b(n))$ *as* $n \to \infty$,

$$b(n)^{-\frac{1}{2}}[nb(n) \int [f_n(x) - f(x)]^2 a(x)\, dx - \int f(x)a(x)\, dx \int w^2(z)\, dz]$$

is asymptotically normally distributed with mean zero and variance

$$2 \int (\int w(x + y)w(x)\, dx)^2\, dy \int a^2(x)f^2(x)\, dx$$

as $n \to \infty$.

The basic technique in obtaining the results is that of approximating the normalized and centered sample distribution function by an appropriate Brownian motion process on a convenient probability space by using a Skorohod-like imbedding due to Brillinger and Breiman. The details of this approximation and remarks on approximation of other functionals are given in Section 2. The asymptotic theory of the maximal deviation and that of quadratic deviations are developed in Sections 3 and 4 respectively. Some computations on the power of these procedures are also carried out. In particular, we show that a goodness-of-fit test based on a quadratic functional is strictly better than the χ^2 test. There is also an appendix on the asymptotic distribution of the maximal deviation for a type of nonstationary Gaussian process.

2. Approximations. As has been indicated in the introduction our technique is to consider the statistics of interest as functionals of certain stochastic processes on the interval [0, 1] and then to substitute Gaussian processes with the same covariance structure for the latter where possible.

It is convenient to introduce $Z_n^0(\cdot)$ given by

$$Z_n^0(t) = n^{\frac{1}{2}}(F_n^*(t) - t)\,, \qquad 0 \leqq t \leqq 1 \tag{2.1}$$

where $F_n^* = F_n(F^{-1})$ is the empirical distribution of $F(X_1), \cdots, F(X_n)$. This will be approximated by $Z^0(\cdot)$, the Brownian bridge, given by

$$Z^0(t) = Z(t) - tZ(1) \tag{2.2}$$

where Z is a standard Wiener process on [0, 1].

The process $[nb(n)f^{-1}(t)]^{\frac{1}{2}}(f_n(\cdot) - E(f_n(\cdot)))$ is central to our discussion. It can be written as

$$Y_n(t) = b^{-\frac{1}{2}}(n)f^{-\frac{1}{2}}(t) \int_{-\infty}^{\infty} w\left(\frac{t - s}{b(n)}\right) dZ_n^0(F(s))\,. \tag{2.3}$$

Approximations ${}_0Y_n$ and ${}_1Y_n$ to this process are obtained by substituting $Z^0(F(\cdot))$ and $Z(F(\cdot))$ respectively for the random measure in (2.3). The resulting processes are well defined, at least if $\int_{-\infty}^{\infty} w^2(t)\, dF(t) < \infty$.

Two other processes which also arise naturally are given by

$$ {}_2Y_n(t) = [b(n)f(t)]^{-\frac{1}{2}} \int w\left(\frac{t - s}{b(n)}\right) (f(s))^{\frac{1}{2}}\, dZ(s) \tag{2.4}$$

and

$$(2.5)\qquad {}_3Y_n(t) = [b(n)]^{-\frac{1}{2}} \int w\left(\frac{t-s}{b(n)}\right) dZ(s)$$

where Z is a two-sided Wiener process on $(-\infty, \infty)$ (dZ is Wiener measure). The process ${}_3Y_n$ is well defined if $\int w^2(t)\,dt < \infty$, and all the integrals with respect to $dZ^0(F(\cdot))$, $dZ(F(\cdot))$, $dZ(\cdot)$, $dZ^0(\cdot)$ are taken in the L^2 sense (cf. Doob [6] page 426). For convenience, suppose all our processes are realized as random elements taking their values in the space $D[0, 1]$ (cf. [3]). For $x \in D[0, 1]$ let $||x|| = \sup\{|x(t)| : 0 \leqq t \leqq 1\}$. Our approximations rest on the following theorem of Brillinger (1969). (A similar argument appeared simultaneously in Breiman 1969).)

THEOREM. *There exists a probability space* (Ω, A, P) *on which one can construct versions of* Z_n^0 *and* Z *such that*

$$(2.6)\qquad ||Z_n^0 - Z^0|| = O_p(n^{-\frac{1}{4}}(\log n)^{\frac{1}{2}}(\log\log n)^{\frac{1}{4}})\,.$$

From this we can derive

PROPOSITION 2.1. *If the processes* Z_n^0, Z^0 *are constructed as above and* A1 *and* A2 *hold, then*

$$(2.7)\qquad ||Y_n - {}_0Y_n|| = O_p(b^{-\frac{1}{2}}(n)n^{-\frac{1}{4}}(\log n)^{\frac{1}{2}}(\log\log n)^{\frac{1}{4}})\,.$$

PROOF. Write, using A1,

$$(2.8)\qquad \begin{aligned} Y_n(q) = [b(n)f(q)]^{-\frac{1}{2}}\{&-w(A)Z_n^0(F(q - Ab(n))) \\ &+ w(-A)Z_n^0(F(q + Ab(n)))\} \\ &+ b^{-\frac{3}{2}}(n)f^{-\frac{1}{2}}(q) \textstyle\int_{-\infty}^{\infty} Z_n^0(F(s))w'\left(\frac{q-s}{b(n)}\right) ds\,. \end{aligned}$$

(The first two terms inside the curly brackets are taken to be 0 in the event A1(b) holds but A1(a) does not.) A similar representation is valid for ${}_0Y_n$ and (2.7) follows.

PROPOSITION 2.2. *If* A2 *holds then*

$$(2.9)\qquad ||{}_0Y_n - {}_1Y_n|| = O_p(b^{\frac{1}{2}}(n))\,.$$

If A2 *and* A3 *hold then*

$$(2.10)\qquad ||{}_2Y_n - {}_3Y_n|| = O_p(b^{\frac{1}{2}}(n))\,.$$

PROOF. From the representation (2.2),

$$(2.11)\qquad \begin{aligned} |{}_0Y_n(q) - {}_1Y_n(q)| &= |Z(1)|[b(n)f(q)]^{-\frac{1}{2}} \\ \int w\left(\frac{q-s}{b(n)}\right) f(s)\,ds &= |Z(1)|O(b^{\frac{1}{2}}(n))\,. \end{aligned}$$

Applying (2.8) and its analogues, if A1(a) holds,

$$
\begin{aligned}
&|{}_2Y_n(q) - {}_3Y_n(q)| \\
&\quad \leqq b^{-\frac{1}{2}}(n)\{[|Z(Ab(n) + q)||[f(Ab(n) + q)/f(q)]^{\frac{1}{2}} - 1| \\
&\quad\quad + |Z(-Ab(n) + q)||[f(-Ab(n) + q)/f(q)]^{\frac{1}{2}} - 1|] \sup_t |w(t)| \\
&\quad\quad + \int |Z(sb(n) + q)||[f(q + sb(n))/f(q)]^{\frac{1}{2}} - 1||w'(s)|\, ds \\
&\quad\quad + \tfrac{1}{2}(b(n)) \int |Z(sb(n) + q)||f'(q + sb(n))|[f(q)f(q + sb(n))]^{-\frac{1}{2}}|w(s)|\, ds\} \\
&\quad = O_p(b^{\frac{1}{2}}(n))
\end{aligned}
\tag{2.12}
$$

by using A3 and the law of the iterated logarithm for the Wiener process. If A1(b) holds the first two terms vanish and the same argument applies.

To apply these propositions we make the elementary

REMARK. If $\{g_n\}$ is a sequence of functionals on $D[0, 1]$ satisfying Lipschitz conditions such that

$$|g_n(x) - g_n(y)| \leqq M_n||x - y|| \tag{2.13}$$

and A_n, B_n are stochastic processes realizable in D such that $||A_n - B_n|| = o_p(1/M_n)$, then $g_n(A_n)$ converges in law if and only if $g_n(B_n)$ does, and to the same limit.

We shall apply this proposition in the next two sections to the functionals

$$\text{I} \qquad (2|\log b(n)|)^{\frac{1}{2}}\left[\max\left\{\frac{|Y_n(t)|}{(\lambda(w))^{\frac{1}{2}}} : 0 \leqq t \leqq 1\right\} - B([b(n)]^{-1})\right]$$

where B is defined in Theorem A1 and,

$$\text{II} \qquad b^{-\frac{1}{2}}(n)\left[\int_{-\infty}^{\infty} Y_n^2(t)f(t)a(t)\, dt - \int_{-\infty}^{\infty} w^2(t)\, dt\right]$$

where a is an integrable weight function. Evidently, since ${}_1Y_n$ and ${}_2Y_n$ have the same joint laws, we can substitute ${}_3Y_n$ for Y_n in I if A1—A3 hold and

$$o\left(\frac{b(n)}{|\log b(n)|}\right) = n^{-\frac{1}{3}} \log n(\log\log n)^{\frac{2}{3}} \tag{2.14}$$

and ${}_0Y_n$ can be substituted for Y_n in II if A1 and A2 hold and,

$$o(b(n)) = n^{-\frac{1}{3}}(\log n)^{\frac{2}{3}}(\log\log n)^{\frac{1}{3}}\,. \tag{2.15}$$

Although we do not pursue this it is clear that the same technique can be applied to other functionals, e.g., a normalized version of the total time in $[0, 1]$ spent by Y_n above a high level (cf. Berman (1971) [2]).

3. The maximum absolute deviation. The first measure of global deviation that we consider is $\tilde{M}_n = \max\{|Y_n(t)| : 0 \leqq t \leqq 1\}$. (There is no loss in considering $[0, 1]$ rather than any other interval on which the density is bounded away from 0 and ∞.) The statistical interest of this functional is twofold as

(i) A convenient way of getting a confidence band for f.
(ii) A test statistic for the hypothesis $H: f = f_0$.

Under (ii) we shall also consider the possibility of testing composite hypotheses, for example, that f is Gaussian. The asymptotic theorem we need to discuss (i), and the behavior of (ii) under the null hypothesis is a consequence of our remarks in Section 2 and Theorem A1 of the appendix.

THEOREM 3.1. *Let w satisfy assumptions* A1—A3 *and*

$$b(n) = n^{-\delta}, \qquad 0 < \delta < \tfrac{1}{2}.$$

Then,

$$P\left[(2\delta \log n)^{\frac{1}{2}} \left(\frac{\tilde{M}_n}{(\lambda(w))^{\frac{1}{2}}} - d_n\right) < x\right] \to e^{-2e^{-x}}, \tag{3.1}$$

where

$$\lambda(w) = \int w^2(t)\, dt \tag{3.2}$$

and

$$d_n = (2\delta \log n)^{\frac{1}{2}} + \frac{1}{(2\delta \log n)^{\frac{1}{2}}} \left\{\log \frac{K_1(w)}{\pi^{\frac{1}{2}}} - \tfrac{1}{2}[\log \delta + \log \log n]\right\} \tag{3.3}$$

where

$$K_1(w) = \frac{w^2(A) + w^2(-A)}{2} \Big/ \lambda(w),$$

if $K_1(w) > 0$, and otherwise

$$d_n = (2\delta \log n)^{\frac{1}{2}} + \frac{1}{(2\delta \log n)^{\frac{1}{2}}} \left[\log \frac{1}{\pi} \frac{K_2(w)}{2}\right]$$

where

$$K_2(w) = \tfrac{1}{2}[\int [w'(t)]^2\, dt]/\lambda(w).$$

REMARK 1. The natural weight function $w(t) = \frac{1}{2}$, $|t| \leqq 1$, $= 0$ otherwise, falls under the first case, while the "optimal" weight function of Epanechnikov (1969) $w(t) = 3/(4(5)^{\frac{1}{2}})(1 - (v^2/5))$ if $|v| \leqq 5^{\frac{1}{2}}$, $= 0$ otherwise, falls under the second.

REMARK 2. A similar result holds if one considers the maximum deviation (rather than absolute deviation) of a density function estimate as in Rosenblatt (1971). However, since one-sided deviations for density functions are unnatural the present result seems more interesting.

REMARK 3. The techniques of proof of this result may readily be adapted to prove limit theorems such as that of Woodroofe (1967) for the maximum deviation observed at an increasing finite number of points.

PROOF. It follows from Propositions 2.1 and 2.2 and the following remark that the limiting behavior of $(2\delta \log n)^{\frac{1}{2}}[(\tilde{M}_n/(\lambda(w))^{\frac{1}{2}}) - d_n]$ is the same as that of $(2 \log b(n))^{\frac{1}{2}}(\max \{|{}_2Y_n(t)|/(\lambda(w))^{\frac{1}{2}} : 0 \leqq t \leqq 1\} - d_n)$. By the similarity transform for the Wiener process, the law

$$L({}_3Y_n(t) : 0 \leqq t \leqq 1) = L\left(\int w\left(\frac{t}{b(n)} - s\right) dZ(s) : 0 \leqq t \leqq 1\right). \tag{3.4}$$

Since $1/(\lambda(w))^{\frac{1}{2}} \int w(t-s)\,dZ(s)$ is a stationary Gaussian process with mean 0 and covariance

$$r(t) = \frac{\int w(s+t)w(s)\,ds}{\lambda(w)}, \tag{3.5}$$

Theorem 3.1 follows from Corollary A.1 provided we show that r satisfies condition (v) and (vi) of Theorem A1 with $\alpha = 1, 2$. That (v) is satisfied is equivalent to Theorem B1. Moreover,

$$\int r^2(t)\,dt = \frac{1}{2\pi}\int |\hat{r}(t)|^2\,dt = \frac{1}{2\pi\lambda^2(w)}\int |\hat{w}(t)|^4\,dt \tag{3.6}$$

where $\hat{}$ denotes Fourier transformation. Since w is integrable and bounded $\hat{w}$ is square integrable and bounded and (vi) must hold.

Applications. (i) To obtain a confidence band for f that is simple and explicit it is natural to consider δ such that $E(f_n)$ can be replaced by f. This is true if $\delta > \frac{1}{5}$ and A4 holds. Then,

$$\frac{1}{b(n)}\int w\left(\frac{t-s}{b(n)}\right)f(s)\,ds = f(t) + O(b^2(n)) \tag{3.7}$$

with 0 independent of t. If we now define $Y_n{}^*$ by replacing $E(f_n(t))$ with $f(t)$ we conclude that

$$||Y_n - Y_n{}^*|| = O([nb^5(n)]^{\frac{1}{2}})\,. \tag{3.8}$$

Using the usual approximations we conclude that $\max\{|Y_n{}^*(t)| : 0 \leqq t \leqq 1\}$ behaves like $\tilde{M}_n$ if A4 holds and $\delta > \frac{1}{5}$. In this case inverting as usual we obtain the confidence band

$$\begin{aligned} f &\leqq f_n + \left(\frac{f_n}{nb(n)}\right)^{\frac{1}{2}} c(\alpha)\left(1 + \frac{c^2(\alpha)}{4nb(n)f_n}\right)^{\frac{1}{2}} + \frac{c^2(\alpha)}{2nb(n)} \\ f &\geqq f_n - \left(\frac{f_n}{nb(n)}\right)^{\frac{1}{2}} c(\alpha)\left(1 + \frac{c^2(\alpha)}{4nb(n)f_n}\right)^{\frac{1}{2}} + \frac{c^2(\alpha)}{2nb(n)} \end{aligned} \tag{3.9}$$

where $c(\alpha)$ is given by (3.11). A simpler band is obtained if we further substitute f_n for f in the denominator of Y_n. The resulting process $Y_n{}^{**}$ (say) has

$$\begin{aligned} ||Y_n{}^* - Y_n{}^{**}|| &= O_p\left(\frac{||Y_n{}^*||^2}{(nb(n))^{\frac{1}{2}}}\,||f_n{}^{-\frac{1}{2}}||\right) \\ &= O_p\left(\frac{\log n}{(nb(n))^{\frac{1}{2}}}\right) \end{aligned} \tag{3.10}$$

if A1—A4 hold and $\frac{1}{5} < \delta < \frac{1}{2}$. The approximate confidence band obtained by looking at the maximum of $|Y_n{}^{**}|$ is given in the introduction (1.2).

There is no choice of δ which asymptotically makes this simple band as thin as possible, i.e. one should choose δ as small as possible. This of course ignores the obvious—the speed with which bias disappears asymptotically depends on δ as does the speed of convergence to the asymptote. However, for fixed n there is an optimal $\delta(n)$ (depending on α) > 0 which for moderate n and small α may be the right thing to use if the choice of bandwidth is free.

(ii) To test $H: f = f_0$ it is natural to compute $\tilde{M}_n$ with $f = f_0$ and reject for large values of the statistic. According to the theorem to obtain approximate level α we should use as cutoff point,

$$c(\alpha) = -[\log|\log(1-\alpha)| - \log 2] \frac{(\lambda(w))^{\frac{1}{2}}}{(2\delta \log n)^{\frac{1}{2}}} + d_n(\lambda(w))^{\frac{1}{2}} . \tag{3.11}$$

Under some assumptions the same cutoff point may be used for testing composite hypotheses of the form $H: f = f_0(\cdot, \theta)$ where θ is an unknown vector parameter by using $\tilde{M}_n$ with an estimate $\hat{\theta}$ substituted for the unknown parameter θ. We need the following assumption.

A5. The estimate $\hat{\theta}$ is such that if $\theta = \theta_0$, for every θ_0,

$$\begin{aligned} \sup\{|\textstyle\int [f_0(t+sb(n), \hat{\theta}) - f_0(t+sb(n), \theta_0)]w(s)\,ds| : 0 \leqq t \leqq 1\} \\ = o_p([nb(n)\log b(n)]^{-\frac{1}{2}}) \end{aligned} \tag{3.12}$$

and

$$||f_0(\cdot, \theta_0) - f_0(\cdot, \hat{\theta})|| = o_p(|\log b(n)|^{-1}) .$$

Typically for maximum likelihood and method of moments estimates

$$|\hat{\theta} - \theta_0| = O_p(n^{-\frac{1}{2}}) . \tag{3.13}$$

If, moreover, $\theta = (\theta^{(1)}, \cdots, \theta^{(k)})$, $\partial f_0/\partial\theta^{(j)}$ is bounded for θ in a neighborhood of θ_0, all x, and $1 \leqq j \leqq k$, it is easy to see that A5 holds. To see that A5 is the needed assumption again introduce a process $\bar{Y}_n$ with $E_\theta(f_n)$ replaced by $E_{\hat{\theta}}(f_n)$ and $(f(\cdot, \theta))^{\frac{1}{2}}$ replaced in the denominator of Y_n by $(f(\cdot, \hat{\theta}))^{\frac{1}{2}}$. Then

$$||Y_n - \bar{Y}_n|| = o_p([\log b(n)]^{-\frac{1}{2}}) \tag{3.14}$$

and the result follows.

To make local power calculations on the test of the simple hypothesis described above we need to consider the behavior of $\tilde{M}_n$ (calculated under f_0) for a sequence of alternatives of the form,

$$g_n(x) = f_0(x) + \gamma_n \eta(x) + o(\gamma_n) \tag{3.15}$$

where g_n satisfy A2—A3 uniformly in n, $\gamma_n \downarrow 0$ at a suitable rate, and $o(\gamma_n)$ is uniform in x on $[0, 1]$. (Note that η must be continuous on $[0, 1]$.) Denote probabilities calculated under g_n by P_n. Our basic result is,

THEOREM 3.2. *Suppose that g_n are as above. Let w satisfy* A1—A3 *and define $\tilde{M}_n$ in terms of f_0. Let*

$$\gamma_n = n^{-\frac{1}{2}+\delta/2}[2\delta \log n]^{-\frac{1}{2}} .$$

Then,

$$P_n\left[(2\delta \log n)^{\frac{1}{2}} \left(\frac{\tilde{M}_n}{(\lambda(w))^{\frac{1}{2}}} - d_n\right) < x\right] \to \exp[-s(\eta)e^{-x}] \tag{3.16}$$

where

$$s(\eta) = \textstyle\int_0^1 \{\exp[\eta(t)/(f_0(t)\lambda(w))^{\frac{1}{2}}] + \exp[-\eta(t)/(f_0(t)\lambda(w))^{\frac{1}{2}}]\}\,dt . \tag{3.17}$$

This result follows from Theorem A1 quite readily.

One interesting consequence of this formula is that our test is asymptotically strictly unbiased for such alternatives. The reason is that $s(\eta) \geqq 2$ with $s(\eta) > 2$ unless $\eta = 0$ and the family of distributions $e^{-\theta e^{-x}}$ is an exponential family in θ.

Unfortunately these tests are asymptotically inadmissible (have Pitman efficiency 0) when compared to the test based on the quadratic functional of the next section based on the same w and $b(n)$. The reason is that alternatives there may be permitted to come in to f_0 at rate $n^{-\frac{1}{2}+\delta/4}$ rather than $n^{-\frac{1}{2}+\delta/2}$. However, this test for moderate sample sizes and some alternatives may well be preferable.

In analogy to the confidence band situation it would appear that maximum power is achieved by taking δ as small as possible. However, consideration of the approximation arguments suggests that $s(\eta_n)$ is a better measure of the "true shift" than $s(\eta)$ where,

$$\eta_n = (g_n - f_0)(2nb(n) \log b(n))^{\frac{1}{2}} . \tag{3.18}$$

Of course, $s(\eta_n)$ may well be maximized for $\delta > 0$. In all of these questions it would be desirable to have some small sample Monte Carlo explorations.

4. Quadratic functionals. We are interested in the behavior of the functional,

$$T_n = nb(n) \int_{-\infty}^{\infty} [f_n(x) - E(f_n(x))]^2 a(x)\, dx = \int_{-\infty}^{\infty} L_n^2(x) a(x)\, dx , \tag{4.1}$$

where $L_n = f^{\frac{1}{2}} Y_n$ and a is integrable. We have already remarked that if A1 and A2 hold and (say) $b_n = n^{-\delta}$, $\delta < \frac{1}{4}$, then.

$$|T_n - \int {}_0L_n^2(x) a(x)\, dx| = o(b^{\frac{1}{2}}(n)) . \tag{4.2}$$

Moreover, if a is bounded as well as integrable and w and f are bounded, we can replace ${}_0L_n$ by ${}_1L_n = f^{\frac{1}{2}}{}_1Y_n$ and hence by ${}_2L_n = f^{\frac{1}{2}}{}_2Y_n$. To see this note that,

$$\begin{aligned} &|\int ({}_1L_n^2(x) - {}_0L_n^2(x)) a(x)\, dx| \\ &\quad = \left| \int \frac{1}{b(n)} \left\{ \left(Z(1) \int w\left(\frac{t-s}{b(n)}\right) f(s)\, ds \right)^2 \right.\right. \\ &\qquad \left.\left. - 2Z(1) \int w\left(\frac{t-s}{b(n)}\right) dZ(F(s)) \int w\left(\frac{t-s}{b(n)}\right) f(s)\, ds \right\} a(t)\, dt \right| \\ &\quad \leqq Z(1)^2 b(n) \sup_x |f(x)| \int |a(t)\, dt| \\ &\qquad + 2|Z(1)|b(n)| \int (\int w(y) c(s + b(n)y) a(s + b(n)y)\, dy)\, dZ(F(s))| \end{aligned} \tag{4.3}$$

where

$$c(t) = \int w(y) f(t - b(n)y)\, dy .$$

But,

$$\begin{aligned} &E(\int (\int w(y) c(s + b(n)y) a(s + b(n)y)\, dy)\, dZ(F(s)))^2 \\ &\qquad = \int (\int w(y) c(s + b(n)y) a(s + b(n)y)\, dy)^2\, dF(s) \end{aligned} \tag{4.4}$$

is bounded.

By (4.3) and (4.4),

$$|T_n - \int {}_2L_n^2(x) a(x)\, dx| = O_p(b(n)) . \tag{4.5}$$

(The infinite range poses no problem since we are approximating L_n rather than the normalized Y_n.)

The following lemma lets us determine the characteristic function of a quadratic functional

$$Z = \int Y(x)^2 a(x)\, dx \tag{4.6}$$

of a Gaussian process $Y(x)$ under appropriate conditions.

LEMMA 4.1. *Let $Y(x)$, $EY(x) \equiv 0$, be a Gaussian process with bounded, uniformly continuous covariance function $r(x, y)$. If $a(x)$ is a piecewise smooth integrable function, the quadratic functional* (4.6) *has characteristic function formally given by*

$$E(e^{itZ}) = \exp\{\textstyle\sum_{k=1}^{\infty} 2^{k-1}(it)^k c_k/k\} \tag{4.7}$$

with

$$c_k = \int \cdots \int r(x_1, x_2)r(x_2, x_3) \cdots r(x_k, x_1)a(x_1)a(x_2) \cdots a(x_k)\, dx_1 \cdots dx_k\,.$$

The representation (4.6) is valid for $|t| < 1/2M$ where $M = ||r|| \int |a(t)|\, dt$. The quantities $(k-1)!\, 2^{k-1}c_k$ are of course the cumulants of (4.6).

The lemma is obtained by considering the form

$$\textstyle\sum_{j=1}^{n} \bar{Y}_j^2 a_j \tag{4.8}$$

in jointly Gaussian random variables $\bar{Y}_j$, $E\bar{Y}_j \equiv 0$ with the a_j's constants. Let R be the covariance matrix of the $\bar{Y}_j$'s with A the diagonal matrix with diagonal entries a_j. The characteristic function of (4.8) is then

$$|1 - 2itRA|^{-\frac{1}{2}} = \textstyle\prod_{j=1}^{n} (1 - 2\lambda_j it)^{-\frac{1}{2}} = \exp\{\sum_{k=1}^{\infty} 2^{k-1}(it)^k \operatorname{tr}(RA)^k/k\}\,,$$

at least if $|t| < 1/2 \operatorname{tr}(RA)$.
Here $\operatorname{tr}(M)$ denotes the trace of M, $|M|$ its determinant and $\lambda_1, \cdots, \lambda_n$ are the eigenvalues of RA. Lemma 4.1 is then obtained by going through an appropriate limiting operation.

The covariance function of the Gaussian process ${}_2L_n(x)$ can be written

$$\begin{aligned} r(x, y) &= \int w(z)w(\alpha + z)f(x - b(n)z)\, dz \\ &= f(x) \int w(z)w(\alpha + z)\, dz + O(b(n)) \end{aligned} \tag{4.9}$$

where

$$\alpha = (y - x)/(b(n))$$

and $O(b(n))$ is independent of x if f is bounded and has a uniformly bounded derivative and $w^2(z)(1 + |z|)$ is integrable. Then

$$E(\textstyle\int {}_2L_n(x)^2 a(x)\, dx) = \int f(x)a(x)\, dx \int w(z)^2\, dz + O(b(n))\,. \tag{4.10}$$

Similarly if a is bounded as well as integrable and w is bounded and f is as above, the variance of $\int {}_2L_n^2(x)a(x)\, dx$ is $2b(n) \int [w * \bar{w}(u)]^2\, du \int a^2(x)f^2(x)\, dx$ to first order as $n \to \infty$, where $\bar{w}(t) = w(-t)$ and $*$ denotes convolution. A similar argument shows that under the same conditions the kth cumulant of $\int {}_2L_n^2(x)a(x)\, dx$ equals to first order $(k-1)!\, 2^{k-1}b^{k-1}(n)[w * \bar{w}]^{(k)}(0) \int a^k(x)f^k(x)\, dx$ as $n \to \infty$ where the superscript (k) indicates that $w * \bar{w}$ is convoluted with itself k times. As a result we have the following theorem which actually holds under the weaker assumptions indicated above.

THEOREM 4.1. *Let* A1—A3 *hold and suppose that a is integrable piecewise continuous and bounded. Suppose moreover that* (2.16) *holds. Then* $b^{-\frac{1}{2}}(n)(T_n - (\int f(x)a(x)\,dx)\int w^2(z)\,dz)$ *is asymptotically normally distributed with mean* 0 *and variance* $2(w * \bar{w})^{(2)}(0)\int a^2(x)f^2(x)\,dx$ *as* $n \to \infty$.

A particular case of interest for the application of the theorem is that in which as in Section 3, $a(x)$ vanishes off an interval, say [0, 1], and one sets $a(x) = f(x)^{-1}$ on [0, 1]. In this case under A1—A3, T_n is asymptotically Gaussian with mean $\int w^2(z)\,dz$ and variance $2b(n)(w * \bar{w})^{(2)}(0)$.

The statistic

$$\tilde{T}_n = nb(n)\int [f_n(x) - f(x)]^2 a(x)\,dx \tag{4.11}$$

is probably of greater interest than that considered in Theorem 4.1. However, let us expand $\tilde{T}_n$ in the form

$$\begin{aligned} nb(n)\{\int [f_n(x) - Ef_n(x)]^2 a(x)\,dx \\ + 2\int [f_n(x) - Ef_n(x)][Ef_n(x) - f(x)]a(x)\,dx \\ + \int [Ef_n(x) - f(x)]^2 a(x)\,dx\}. \end{aligned} \tag{4.12}$$

Let w be positive and symmetric about zero with

$$c = \int w(u)u^2\,du < \infty\,. \tag{4.13}$$

Then if $n^{-1} = O(b(n))$, $b(n) \to 0$ as $n \to \infty$, A1 holds and f has a continuous bounded second derivative, the second term of (4.12) may, by the usual approximation arguments, be shown to be asymptotically normal with mean zero and variance

$$n^{-1}b(n)^4c^2\int f''(x)^2a(x)^2f(x)\,dx \tag{4.14}$$

to the first order. Also, under the same conditions, the last term of (4.12) can be shown to be

$$b(n)^4c^2\int f''(x)^2a(x)\,dx \tag{4.15}$$

to the first order. Then $[b(n)]^{-\frac{1}{2}}[\tilde{T}_n - T_n] = o_p(1)$ if and only if $b(n) = o(n^{-\frac{2}{9}})$. (The term (4.14) is then negligible.) The theorem quoted in the introduction follows.

APPLICATIONS. An explicit confidence band is hard to obtain from Theorem 4.1 and the theorem of the introduction. However we can test $H: f = f_0$ at (approximate) level α by calculating T_n for $f = f_0$ and rejecting when $T_n \geqq d(\alpha)$ where by Theorem 4.1

$$\begin{aligned} d(\alpha) = [\int f_0(x)a(x)\,dx][\int w^2(z)\,dz] \\ + b^{\frac{1}{2}}(n)\Phi^{-1}(1-\alpha)/[2(w * \bar{w})^{(2)}(0)\int a^2(x)f_0^2(x)\,dx]^{\frac{1}{2}}\,. \end{aligned} \tag{4.16}$$

As in Section 3 it is easy to see that in testing $H: f = f_0(\cdot, \theta)$ where θ is an unknown vector parameter we may use T_n with f replaced by $f_0(\cdot, \hat{\theta})$ and $d(\alpha)$ with f_0 replaced by $f_0(\cdot, \hat{\theta})$, provided that A6 below holds.

A6. For each θ_0, $(\partial^2 f(x, \theta)/\partial\theta^{(i)}\partial\theta^{(j)})$ is bounded in absolute value for all θ in a neighborhood of θ_0 and all x, i, j. Moreover, if θ_0 is true,

$$|\hat{\theta} - \theta_0| = o_p([nb(n)]^{-\frac{1}{2}}) . \tag{4.17}$$

To see this, taking $k = 1$ for simplicity, expand as in (4.12) and note that it suffices to show that

$$\int [f_n(x) - E_{\theta_0}(f_n(x))][E_{\theta_0}(f_n(x)) - E_{\hat{\theta}}(f_n(x))]a(x)\, dx = o_p([nb^{\frac{1}{2}}(n)]^{-1}) \tag{4.18}$$

and

$$\int [E_{\theta_0}(f_n(x)) - E_{\hat{\theta}}(f_n(x))]^2 a(x)\, dx = o_p([nb^{\frac{1}{2}}(n)]^{-1}) . \tag{4.19}$$

Taylor expanding the integral in (4.18) about θ_0 we obtain a first term

$$(\hat{\theta} - \theta_0) \int [f_n(x) - E_{\theta_0}(f_n(x))]\left[\int \frac{\partial f(x + b(n)z, \theta)}{\partial\theta}\bigg|_{\theta=\theta_0} w(z)\, dz\right] a(x)\, dx$$

which is $O_p(|\hat{\theta} - \theta_0|n^{-\frac{1}{2}})$, and a second term which is $O_p([nb(n)]^{-\frac{1}{2}}(\hat{\theta} - \theta_0)^2)$, and (4.18) follows. A similar argument yields (4.19).

To make local power calculations we again suppose g_n is as in (3.15) with g_n satisfying A2—A3 uniformly in n and $o(\gamma_n)$ uniform in x and η is bounded.

THEOREM 4.2. *Let g_n be as above, w satisfy* A1—A4, *a be integrable piecewise continuous and bounded, $b(n) = n^{-\delta}$, $\delta < \frac{1}{4}$, $\gamma_n = n^{-\frac{1}{2}+\delta/4}$. Define T_n in terms of f_0. Then,*

$$b^{-\frac{1}{2}}(n)(T_n - [\int f_0(x)a(x)\, dx] \int w^2(z)\, dz) \tag{4.20}$$

is asymptotically normally distributed with mean $\int \eta^2(x)a(x)\, dx$ and variance

$$2(w * \tilde{w})^{(2)}(0) \int a^2(x)f_0^2(x)\, dx .$$

The proof is straightforward. As in Section 3 it follows that the test which rejects when T_n is $\geqq d(\alpha)$ is locally strictly unbiased if $a(x) > 0$ for all x.

Also as before the asymptotics lead to choosing δ as large as possible and again this conclusion is shaken if one uses the better approximation to the asymptotic mean, $\int \eta_n^2(x)a(x)\, dx$ where

$$\eta_n(x) = \int w(z)[g_n(x + b(n)z) - f_0(x + b(n)z)]\, dz . \tag{4.21}$$

It is also clear that for fixed δ we can let $\lambda_n \to 0$ more quickly than for the sup functional and still get power. Thus the Pitman efficiency of the T_n test to the $\tilde{M}_n$ test for the same δ is ∞.

Suppose that f_0 is the uniform density on [0, 1] an effect we can always achieve by applying the probability integral transformation to our observations before making the test. Let $a(x) = 1$ on [0, 1] and 0 otherwise, w be the uniform density on $[-\frac{1}{2}, \frac{1}{2}]$. Neglecting fringe effects we may then write

$$T_n = \int_0^1 \frac{(N[t - \frac{1}{2}b(n), t + \frac{1}{2}b(n)] - b(n))^2}{nb(n)} dt \tag{4.22}$$

where $N[x, y]$ is the number of observations falling in the interval $[x, y]$. A related statistic for testing uniformity on the circle was considered by Watson in [15]. This is, of course, very similar to the χ^2 statistic for the problem based on the cells $[0, b(n)]$, $[b(n), 2b(n)]$, $\cdots$, $[(K-1)\frac{1}{2}b(n), (K+1)\frac{1}{2}b(n)]$ given by,

$$\chi_n^2 = \sum_{k=1}^{*K} \frac{(N[\frac{1}{2}kb(n) - \frac{1}{2}b(n), \frac{1}{2}kb(n) + \frac{1}{2}b(n)] - nb(n))^2}{nb(n)} \tag{4.23}$$

where $(K+1)\frac{1}{2}b(n) \leqq 1 < (K+2)\frac{1}{2}b(n)$ and $\sum^*$ is a sum over odd index.

Now we can write,

$$\chi_n^2/K = nb(n) \int_0^1 (f_n(t) - E(f_n(t)))^2 \, dA_n(t) \tag{4.24}$$

where A_n places mass $1/K$ at each of the points $\frac{1}{2}b(n)$, $\cdots$, $K\frac{1}{2}b(n)$. It is easy to see that the arguments leading to Theorem 4.2 apply to functionals of this type also and that under the conditions of that theorem, if $b(n) = n^{-\delta}$, $\delta < \frac{1}{2}$, χ_n^2/K is asymptotically normal with the natural parameters $E(\chi_n^2/K)$ and Var (χ_n^2/K).

This result is, of course, known. A rigorous proof under milder conditions but using a different method may be found in Steck (1957). Now

$$\begin{aligned} E\left(\frac{\chi_n^2}{K}\right) &= 1 + \frac{1}{K} \sum_{j=1}^{*K} nb(n)\left(1 - \frac{1}{b(n)} \int_{(j-1)b(n)/2}^{(j+1)b(n)/2} g_n(x)\,dx\right)^2 \\ &= 1 + nb(n)\gamma_n^2 \frac{1}{K} \sum_{j=1}^{*K} \left[\frac{1}{b(n)} \int_{(j-1)b(n)/2}^{(j+1)b(n)/2} \eta(x)\,dx\right]^2 + o(nb(n)\gamma_n^2) \end{aligned} \tag{4.25}$$

$$\operatorname{Var}\left(\frac{\chi_n^2}{K}\right) = \frac{1}{K} \operatorname{Var}\left(\frac{N^2[0, b(n)]}{nb(n)}\right) + o\left(\frac{1}{K}\right) = \frac{2}{K} + o\left(\frac{1}{K}\right). \tag{4.26}$$

Thus if we take $\gamma_n = n^{-\frac{1}{2}+\delta/4}$ as in Theorem 4.2, under g_n the statistics

$$W_n = b^{-\frac{1}{2}}(n)\left(\frac{\chi_n^2}{K} - 1\right) \tag{4.27}$$

have a limiting Gaussian distribution with mean $\int_0^1 \eta^2(x)\,dx$ and variance 2. Under the same circumstances the asymptotic mean of $b^{-\frac{1}{2}}(n)(T_n - 1)$ with T_n given by (4.22) is also $\int_0^1 \eta^2(x)\,dx$ while its asymptotic variance is,

$$2w^{(4)}(0) = 2\int_{-1}^{1} (1 - |t|)^2\,dt = \tfrac{4}{3}. \tag{4.28}$$

The Pitman efficiency of the tests based on T_n to those based on W_n is thus by the usual calculations,

$$e(T_n, W_n) = (\tfrac{3}{2})^{\frac{1}{2}-\delta} \tag{4.29}$$

and thus at least $(\frac{3}{2})^{\frac{1}{2}} = 1.217$ on the range $\delta > 0$. For the Mann-Wald (1942) prescription $\delta = \frac{2}{5}$ we get an efficiency of 1.292.

Although as we have seen these asymptotic calculations are to be taken with a grain of salt we feel that the procedure T_n has promise as a competitor to the χ^2 test, at least for moderate sample sizes.

Acknowledgment. We are grateful to S. Berman and J. Pickands, III for providing us with preprints of some of the papers cited in the list of references.

5. Appendix A. On the extrema of some nonstationary Gaussian processes. Let $Y_T(\cdot)$ be a sequence of separable Gaussian processes with mean $\mu_T(\cdot)$ such that $Y_T(\cdot) - \mu_T(\cdot)$ is stationary. Let $r(\cdot)$ be the covariance function of Y_T,

$$M_T = \max\{Y_T(t): 0 \leqq t \leqq T\}, \qquad m_T = \min\{Y_T(t): 0 \leqq t \leqq T\}.$$

Let $b_T(t) = \mu_T(t)(2\log T)^{\frac{1}{2}}$.

THEOREM A1. *Suppose that,*

(i) $b_T(t)$ *is uniformly bounded in* t *and* T *on* $[0, T]$ *as* $T \to \infty$.

(ii) $b_T(t) \to b(t)$ *uniformly on* $[0, T]$ *as* $T \to \infty$.

(iii) $T^{-1}\lambda[t: b(t) \leqq x, 0 \leqq t \leqq T] \to \eta(x)$ *the* cdf *of a probability measure as* $T \to \infty$. *(λ as usual denotes Lebesgue measure.)*

(iv) $b(\cdot)$ *is uniformly continuous on* R.

(v) $r(t) = 1 - C|t|^\alpha + o(|t|^\alpha)$, $0 < \alpha \leqq 2$, *as* $t \to \infty$.

(vi) $\int_0^\infty r^2(t)\,dt < \infty$.

Let

$$B(t) = (2\log t)^{\frac{1}{2}} + \frac{1}{(2\log t)^{\frac{1}{2}}} \times \left\{\left(\frac{1}{\alpha} - \frac{1}{2}\right)\log\log t + \log(2\pi)^{-\frac{1}{2}}(C^{1/\alpha}H_\alpha 2^{(2-\alpha)/2\alpha})\right\}$$

where

$$H_\alpha = \lim_{T\to\infty}\frac{1}{T}\int_0^\infty e^s P[\sup_{0\leq t\leq T} Y(t) > s]\,ds$$

and Y *is a Gaussian process with,*

$$E(Y(t)) = -|t|^\alpha, \qquad \operatorname{Cov}(Y(t_1), Y(t_2)) = |t_1|^\alpha + |t_2|^\alpha - |t_1 - t_2|^\alpha. \tag{A.1}$$

Then,

$$U_T = (2\log T)^{\frac{1}{2}}(M_T - B(T)) \qquad \textit{and} \qquad V_T = -(2\log T)^{\frac{1}{2}}(m_T + B(T))$$

are asymptotically independent with,

$$P[U_T < z] \to e^{-\lambda_1 e^{-z}}, \qquad P[V_T < z] \to e^{-\lambda_2 e^{-z}}; \tag{A.2}$$

where,

$$\lambda_1 = \int e^z\,d\eta(z), \qquad \lambda_2 = \int e^{-z}\,d\eta(z). \tag{A.3}$$

An immediate consequence of Theorem A1 is, COROLLARY A1. *If* $\tilde{M}_T = \max\{|Y_T(t)| : 0 \leqq t \leqq T\}$ *then under the conditions of the theorem,*

$$P[(2\log T)^{\frac{1}{2}}(\tilde{M}_T - B(T)) < x] \to \exp[-(\lambda_1 + \lambda_2)e^{-x}]\,. \tag{A.4}$$

Note. $\lambda_1 + \lambda_2 \geqq 2$ with strict inequality unless η concentrates at 0.

COROLLARY A2. *Let* $Y_0(t) - \mu(t)$ *be a stationary mean* 0 *Gaussian process with covariance function* $r(t)$ *satisfying the conditions of the theorem. Suppose that* $b(t) = (2\log(t+2))^{\frac{1}{2}}\mu(t)$ *is a bounded uniformly continuous function of* t *and that* $b(\cdot)$ *satisfies condition* (iii) *of the theorem. Then,*

$$P[(2\log T)^{\frac{1}{2}}(\max\{Y_0(s) : 0 \leqq s \leqq T\} - B(T)) < x] \to e^{-\lambda_1 e^{-x}}\,. \tag{A.5}$$

Similar assertions hold about the independence of maximum and minimum and the asymptotic distribution of the minimum.

This corollary may be viewed as complementing Theorem 4.1 of Qualls and Watanabe (1971) which deals with the extrema of a mean 0 process whose covariance function is asymptotically locally approximated by that of a stationary process while we deal with a process which is stationary when centered and asymptotically stationary.

The constants H_1 and H_2 are the only ones known explicitly. They are given by $H_1 = 1$, $H_2 = \pi^{-\frac{1}{2}}$ (cf. [11]).

PROOF OF COROLLARY A2. Define,

$$\begin{aligned} Y_T(t) &= Y_0(t) \qquad \text{on} \quad [\varepsilon(T), T] \\ &= Y_0(t) + ([\log(t+2)/\log(T+2)]^{\frac{1}{2}} - 1)\mu(t) \quad \text{otherwise} \end{aligned} \tag{A.6}$$

where $\varepsilon(T) = o(T)$, $\log\varepsilon(T) \sim \log T$. Evidently, $(2\log T)^{\frac{1}{2}}E(Y_T(t)) \to b(t)$ uniformly and

$$\begin{aligned} &\left| P\left[\max\{Y_T(s) : 0 \leqq s \leqq T\} < \frac{x}{(2\log T)^{\frac{1}{2}}} + B(T)\right] \right. \\ &\qquad \left. - P\left[\max\{Y_0(s) : 0 \leqq s \leqq T\} < \frac{x}{(2\log T)^{\frac{1}{2}}} + B(T)\right] \right| \\ &\qquad \leqq 2P\left[\max\{Y_0(s) - E(Y_0(s)) : 0 \leqq s \leqq \varepsilon(T)\} \geqq \frac{x}{(2\log T)^{\frac{1}{2}}} \right. \\ &\qquad\qquad \left. - K + B(T)\right] \end{aligned} \tag{A.7}$$

where $K = \max\{\mu(t) : 0 \leqq t \leqq \varepsilon(t)\}$. Since $B(\varepsilon(T)) - B(T) \to -\infty$ the term on the right of (A.7) tends to 0 by the theorem. □

PROOF OF THEOREM A1. The theorem is argued much as Theorem 3.1 of Pickands (1969). We refer the reader to this paper and Berman (1971) for the details of the argument.

LEMMA A1. *Let $\psi(x) = \phi(x)/x$ where ϕ is the standard normal density. Let $C = 1$, $x = x(T) = B(T) + z_1/(2\log T)^{\frac{1}{2}}$. Then for $a > 0$,*

$$\begin{aligned} &P[\max\{Y_T(t + akx^{-2/\alpha}), 0 \leqq k \leqq n\} > x] \\ &\qquad = \psi(x)e^{b(t)}H_\alpha(n, \alpha) + o(\psi(x)) \\ &P[\min\{Y_T(t + kax^{-2/\alpha}), 0 \leqq k \leqq n\} < -x] \\ &\qquad = \psi(x)e^{-b(t)}H_\alpha(n, a) + o(\psi(x)) \end{aligned} \tag{A.8}$$

as $T \to \infty$ uniformly in $0 \leqq t \leqq T$ where

$$H_\alpha(n, a) = \int_{-\infty}^{\infty} e^s P[\max\{Y(ka) : 0 \leqq k \leqq n\} > s]\, ds\,. \tag{A.9}$$

Moreover, if $y = y(T) = B(T) + z_2/(2\log T)^{\frac{1}{2}}$ then

$$\begin{aligned} &P[\max\{Y_T(t + kax^{-2/\alpha}) : 0 \leqq k \leqq n\} > x, \\ &\qquad \min\{Y_T(t + kax^{-2/\alpha}) : 0 \leqq k \leqq n\} < -y] \\ &\qquad = o(\psi(x)) = o(\psi(y))\,, \end{aligned} \tag{A.10}$$

uniformly in $0 \leqq t \leqq T$. (Throughout, k may take on integer values only.)

PROOF. As in [11] consider the "local" process

$$\tilde{Y}_T(s) = x(Y_T(t + sx^{-2/\alpha}) - \mu_T(t) - x)\,. \tag{A.11}$$

$$\begin{aligned} &P[\max\{Y_T(t + akx^{-2/\alpha}) : 0 \leqq k \leqq n\} < x] \\ &\qquad = \int_{-\infty}^{\infty} \gamma(z) P[\max\{\tilde{Y}_T(ka) : 0 \leqq k \leqq n\} > -x\mu_T(t) \mid \tilde{Y}_T(0) = z]\, dz \end{aligned} \tag{A.12}$$

where γ is the density of $\tilde{Y}_T(0)$,

$$\gamma(z) = \frac{1}{x}\phi\left(x + \frac{z}{x}\right) = \psi(x)\exp[-z - z^2/2x^2]\,. \tag{A.13}$$

It is easy to see using (ii) and (iv) that the finite dimensional conditional distributions of $\tilde{Y}_T(s)$ given $\tilde{Y}_T(0) = z$ converge uniformly in t to those of the process $Y(s) + z$ where Y is given by (A.1). Arguing as in [11] the first part of (A.8) follows since $x\mu_T(t) \to b(t)$ uniformly as required. By considering $-Y_T$ we obtain the second part. To prove (A.10) let A be the event whose probability is being estimated. Then,

$$\begin{aligned} &P\left(A, Y_T(t) > x - \frac{1}{x^{\frac{1}{2}}} + \mu_T(t)\right) \\ &\quad \leqq \int_{-x^{\frac{1}{2}}}^{\infty} \gamma(z) P[\min\{\tilde{Y}_T(ka) : 0 \leqq k \leqq n\} - z \\ &\quad \leqq -z - x(y + x + \mu_T(t)) \mid \tilde{Y}_T(0) = z]\, dz \\ &\quad \leqq \psi(x) \int_{-\infty}^{x^{\frac{1}{2}}} e^z P[\min\{\tilde{Y}_T(ka) + z : 0 \leqq k \leqq n\} \\ &\quad < z - x(y + x + \mu_T(t)) \mid \tilde{Y}_T(0) = -z]\, dz \\ &\quad \leqq \psi(x) \sum_{k=0}^{n} \{\int_{-\infty}^{0} P[\tilde{Y}_T(ka) + z < z \\ &\qquad - x(y + x + \mu_T(t)) \mid \tilde{Y}_T(0) = -z]\, dz \\ &\qquad + x^{\frac{1}{2}} \exp x^{\frac{1}{2}} \max\{P[\tilde{Y}_T(ka) + z \\ &\quad < x^{\frac{1}{2}} - x(y + x + \mu_T(t)) \mid \tilde{Y}_T(0) = -z] : 0 \leqq z \leqq x^{\frac{1}{2}}\}\}]\,. \end{aligned} \tag{A.14}$$

Applying the usual estimate $\Phi(z) \leqq \psi(|z|)$ for $z \leqq 0$ we conclude that the left-hand side of (A.14) is $o(\psi(x))$. Similarly,

$$P\left(A,\, Y_T(t) - \mu_T(t) < -y + \frac{1}{y^{\frac{1}{2}}}\right) = o(\psi(y))\,. \tag{A.15}$$

Finally,

$$\begin{aligned} &P\left(A,\, -y + \frac{1}{y^{\frac{1}{2}}} \leqq Y_T(t) - \mu_T(t) \leqq x - \frac{1}{x^{\frac{1}{2}}}\right) \\ &\quad \leqq \int_{-\infty}^{-z^{\frac{1}{2}}} \gamma(z) P[\max\{\tilde{Y}_T(ka) : 0 \leqq k \leqq n\} > -x\mu_T(t) \mid \tilde{Y}(0) = z]\, dz \\ &\quad \leqq \psi(x) \int_A^\infty e^z P[\max\{\tilde{Y}_T(ka) : 0 \leqq k \leqq n\} > z]\, dz \end{aligned} \tag{A.16}$$

for every $A < \infty$.

The final statement of the lemma follows.

LEMMA A2. *The assertion of Lemma* A1 *remains valid if* $a = 1$, k *is permitted to range over all values in* $[0, n]$ *and* $H_\alpha(n, a)$ *is replaced by*

$$\bar{H}_\alpha(n) = \cdot \int_{-\infty}^{\infty} e^t P[\max\{Y(s) : 0 \leqq s \leqq n\} > t]\, dt \tag{A.17}$$

PROOF. We prove the analogue of (A.8); the other assertions follow similarly. We need to check that uniformly in T,

(a) The conditional distributions of the continuous processes $\tilde{Y}_T(t) - z$ given $\tilde{Y}_T(0) = z$ converge weakly (in the sense of Prohorov) to that of $Y(\cdot)$,

(b) $$P[\max\{\tilde{Y}_T(k) : 0 \leqq k \leqq n\} > x\mu_T(t) \mid \tilde{Y}_T(0) = z] \leqq g(z)$$

where $\int e^{-z} g(z)\, dz < \infty$.

To see that (a) holds it suffices to note that,

$$\operatorname{Var}[(\tilde{Y}_T(s_1) - \tilde{Y}_T(s_2)) \mid \tilde{Y}_T(0) = z] \leqq C|s_1 - s_2|^\alpha \tag{A.18}$$

and then apply Billingsley [3] page 95. To see that (b) is valid use the estimate of Fernique (1970) given below on the tails of $\max\{|\tilde{Y}_T(k)| : 0 \leqq k \leqq n\}$.

LEMMA. *Let* $Z(\cdot)$ *be a Gaussian process on* $(0, 1)$. *Let* a *be such that* $P[||Z|| \leqq a] \geqq \frac{3}{4}$, $P[||Z|| \geqq a] \geqq \frac{1}{4}$. *Then, for* $z \geqq a$

$$P[||Z|| > z] \leqq \exp\left\{-\frac{z^2}{24a^2} \log 3\right\}.$$

LEMMA A3. *Fix* $t > 0$ *such that* $\inf\{s^{-\alpha}(1 - r(s)) : 0 \leqq s \leqq t\} \geqq A(t) > 0$. *Define* x *and* y *as before. Let,*

$$H_\alpha(a) = \lim_{n\to\infty} \frac{H_\alpha(n, a)}{n}\,. \tag{A.19}$$

$$0 < H_\alpha = \lim_{a\to 0} \frac{H_\alpha(a)}{a} = \lim_{n\to\infty} \frac{\bar{H}_\alpha(n)}{n}\,. \tag{A.20}$$

(See the note at the end of the lemma.)

Then,

$$P\left[\max\left\{Y_T(v + kax^{-2/\alpha}) : 0 \leqq k \leqq \left[\frac{x^{2/\alpha}}{a}t\right]\right\} > x\right] = x^{2/\alpha}\psi(x)\frac{H_\alpha(a)}{a}\int_v^{v+t}\exp b(s)\,ds + o(x^{2/\alpha}\psi(x)), \tag{A.21}$$

$$P[\max\{Y_T(v + s) : 0 \leqq s \leqq t\} > x] = x^{2/\alpha}\psi(x)\left[\int_v^{v+t}\exp b(s)\,ds\right]H_\alpha + o(x^{2/\alpha}\psi(x)), \tag{A.22}$$

uniformly in $0 \leqq v \leqq T$. *Similar assertions hold for* $P[\min\{Y_T(v + s) : 0 \leqq s \leqq t\} < -x]$ *with* $-b$ *replacing* b. *Finally,*

$$P[\max\{Y_T(v + s) : 0 \leqq s \leqq t\} > x, \min\{Y_T(v + s) : 0 \leqq s \leqq t\} < -y] = o(x^{2/\alpha}\psi(x)). \tag{A.23}$$

Note. The existence of the limit in (A.19) was first proved in [11]. An incorrect proof of (A.20) was also given. Subsequently, a correct proof was communicated to the author by J. Pickands and another is included in [12]. We provide yet a third in Appendix B.

PROOF. We prove (A.22); (A.21) is argued similarly. Begin by bounding the left-hand side of (A.22) from above by,

$$\sum_{k=0}^{M} P[\max\{Y_T(v + knx^{-2/\alpha} + s) : 0 \leqq s \leqq nx^{-2/\alpha}\} > x] \tag{A.24}$$

where $M = [tx^{2/\alpha}/n]$. By Lemma A2 the expression above is asymptotic to

$$\frac{t\bar{H}_\alpha(n)}{n}x^{2/\alpha}\psi(x)\left[\frac{1}{M+1}\sum_{k=0}^{M}\exp b(v + knx^{-2/\alpha})\right] = \frac{\bar{H}_\alpha(n)}{n}x^{2/\alpha}\psi(x)\left[\int_v^{v+t}\exp b(s)\,ds + o(1)\right] \tag{A.25}$$

since b is assumed uniformly continuous and bounded. On the other hand we can bound from below by the left-hand side of (A.21) which in turn is bounded from below by,

$$\sum_{r=0}^{M_a} P(A_r) - \sum_{0\leq r\neq s\leq M_a} P(A_r A_s) \tag{A.26}$$

where $A_r = [\{\max\{Y_T(v + kax^{-2/\alpha}), rn \leqq k < (r + 1)n\} > x]$, $M_a = [x^{-2/\alpha}t/na]$. If we apply Lemma A.1 to the first term on the right of (A.26) we obtain that,

$$\sum_{r=0}^{M_a} P(A_r) \sim \frac{H_\alpha(n, a)}{na}x^{2/\alpha}\psi(x)\left[\int_v^{v+t}e^{b(s)}\,ds\right]. \tag{A.27}$$

Finally,

$$P(A_r A_s) \leqq P(C_r C_s) \tag{A.28}$$

where

$$C_r = \left[\max\{Y_T(kax^{-2/\alpha} + v) - \mu_T(kax^{-2/\alpha}) : rn \leqq k < (r + 1)n\} > x - \frac{K}{(2\log T)^{\frac{1}{2}}}\right] \tag{A.29}$$

where $K = \sup\{(2\log T)^{\frac{1}{2}}|\mu_T(t)| : 0 \leqq t \leqq T\}$. Now applying Lemma 2.3 of [11] and arguing as in Lemma 2.5 of the same paper we see that,

$$\sum P(C_r C_s) = o(x^{2/\alpha}\psi(x)) . \tag{A.30}$$

Applying (A.20) we see that (A.22) follows. To prove (A.23) it suffices to show that,

$$\begin{aligned} P\{[\max\{Y_T(v+s) : 0 \leqq s \leqq t\} > x] \\ \cup [\min\{Y_T(v+s) : 0 \leqq s \leqq t\} < -y]\} \\ = x^{2/\alpha}\psi(x)H_\alpha \textstyle\int_v^{v+t} [\exp b(\xi)]\, d\xi \\ + y^{2/\alpha}\psi(y)H_\alpha \textstyle\int_v^{v+t} [\exp -b(\xi)]\, d\xi + o(x^{2/\alpha}\psi(x)) . \end{aligned} \tag{A.31}$$

But we can bound the expression on the left of (A.31) from above by

$$P[\max\{Y_T(v+s) : 0 \leqq s \leqq t\} > x] + P[\min\{Y_T(v+s) : 0 \leqq s \leqq t\} < -y]$$

and from below as in (A.26) where we add $A_{M_a+1}, \cdots, A_{2M_a+1}$ with $A_{M_a+j} = \{\min\{Y_T(v + kax^{-2/\alpha}) : (j-1)n \leqq k < jn\}\} < -y\}$. Now by (A.10)

$$\frac{1}{n} \textstyle\sum_{j=0}^{M_a} P(A_j A_{M_a+j+1}) = o(x^{2/\alpha}\psi(x)) . \tag{A.32}$$

Finally, again arguing as for the previous case,

$$\frac{1}{n} \textstyle\sum_{0 \leqq j \neq k \leqq M_a} P(A_j A_k) , \qquad \frac{1}{n} \sum_{1 \leqq j \neq k \leqq M_a+1} P(A_{j+M_a} A_{k+M_a}) \qquad \text{and} \tag{A.33}$$

$$\frac{1}{n} \textstyle\sum_{0 \leqq j \neq k \leqq M_a} P(A_j A_{M_a+k+1}) \quad \text{are all} \quad o(x^{2/\alpha}\psi(x)) . \ \square$$

The rest of the proof goes much as in Berman [1]. Neglecting fringe effects break the interval $[0, T]$ up into $2N$ intervals of which half, $W_1, \cdots, W_N$ are of length t and the others $V_1, \cdots, V_N$ of length ε so that V_i follows W_i which follows V_{i-1}, $i = 2, \cdots, N$. Of course, $N \sim T/(t+\varepsilon)$. Define x and y as in Lemma A1 and note that,

$$x^{2/\alpha}\psi(x)H_\alpha \sim \frac{1}{T} e^{-z_1} . \tag{A.34}$$

Then, by Lemma A3,

$$\begin{aligned} P[\max\{Y_T(\tau) : \tau \in \textstyle\bigcup_{j=1}^N V_j\} \geqq x] &\leqq \textstyle\sum_{j=1}^N P[\max\{Y_T(\tau) : \tau \in V_j\} \geqq x] \\ &\sim [\textstyle\sum_{j=1}^N \int_{V_j} \exp b(s)\, ds] \frac{e^{-z_1}}{T} \\ &= \varepsilon O\left(\frac{N}{T}\right) = \varepsilon O(1) \end{aligned} \tag{A.35}$$

where the O term is independent of ε and the V_j. A similar assertion holds for $\min\{Y_T(\tau) : \tau \in \bigcup_{j=1}^N V_j\}$ and hence we need only show that,

$$\begin{aligned} \lim_{\varepsilon\to 0} \underline{\lim}_{T\to\infty} P[\max\{Y_T(\tau) : \tau \in \textstyle\bigcup_{j=1}^N W_j\} \leqq x, \\ \min\{Y_T(\tau) : \tau \in \textstyle\bigcup_{j=1}^N W_j\} \geqq -y] \\ = \exp -\{\lambda_1 e^{-z_1} + \lambda_2 e^{-z_2}\} , \end{aligned} \tag{A.36}$$

where λ_1, λ_2 are defined in (A.3) and the bars above and below the limit sign indicate lim sup and lim inf respectively. Next choose $a > 0$. If $W_j = [a_j, a_j + t)$, $j = 1, \cdots, N$.

$$\begin{aligned}
&\Big| P[\max\{Y_T(\tau) : \tau \in \textstyle\bigcup_{j=1}^N W_j\} \leqq x] \\
&\quad - P\left[Y_T(a_j + kax^{-2/\alpha}) \leqq x : 0 \leqq k \leqq \left[\frac{tx^{2/\alpha}}{a}\right], 1 \leqq j \leqq N\right]\Big| \\
&\leqq \textstyle\sum_{j=1}^N \Big| P[\max\{Y_T(\tau) : \tau \in W_j\} \leqq x] \\
&\quad - P\left[\max\left\{Y_T(a_j + kax^{-2/\alpha}) : 0 \leqq k \leqq \left[\frac{tx^{2/\alpha}}{a}\right]\right\} \leqq x\right]\Big| \\
&\sim \left[\textstyle\sum_{j=1}^N \int_{W_j} \exp b(s)\, ds\right] x^{2/\alpha}\psi(x)\left[H_\alpha - \frac{H_\alpha(a)}{a}\right] e^{-z_1},
\end{aligned} \tag{A.37}$$

by Lemma A3.

A similar argument holds for $P[\min\{Y_T(\tau) : \tau \in \bigcup_{j=1}^N W_j\} \geqq -y]$ and by simple probability manipulations it follows that to prove the theorem we need only show,

$$\begin{aligned}
&\lim_{a\to 0} \lim_{\varepsilon\to 0} \overline{\underline{\lim}}_T\, P\Big[-y \leqq Y_T(a_j + kax^{-2/\alpha}) \leqq x : 1 \leqq j \leqq N, \\
&\qquad 0 \leqq k \leqq \left[\frac{tx^{2/\alpha}}{a}\right]\Big] \\
&\qquad\qquad = \exp\{-[\lambda_1 e^{-z_1} + \lambda_2 e^{-z_2}]\}.
\end{aligned} \tag{A.38}$$

Now in view of Lemma A3 it is easy to show that,

$$\begin{aligned}
&\underline{\lim}_T \textstyle\sum_{j=1}^N \left(1 - P\left[-y \leqq Y_T(a_j + kax^{-2/\alpha}) \leqq x : 0 \leqq k \leqq \left[\frac{tx^{2/\alpha}}{a}\right]\right]\right) \\
&\quad = \frac{H_\alpha(a)}{aH_\alpha} \lim \frac{1}{T} \textstyle\sum_{j=1}^N \int_{W_j} \{\exp[b(s) - z_1] + \exp -[b(s) + z_2]\}\, ds.
\end{aligned} \tag{A.39}$$

Since, by the boundedness of b, $T^{-1}[\sum_{j=1}^N \int_{W_j} \exp b(s)\, ds - \int_0^T \exp b(s)\, ds] = O(\varepsilon)$ uniformly in T it follows from (A.39) and (A.20) that

$$\begin{aligned}
&\lim_{a\to 0} \lim_{\varepsilon\to 0} \underline{\lim}_T \textstyle\sum_{j=1}^N \Big(1 - P\Big[-y \leqq Y_T(a_j + kax^{-2/\alpha}) \\
&\qquad \leqq x : 0 \leqq k \leqq \left[\frac{tx^{2/\alpha}}{a}\right]\Big]\Big) \\
&\qquad\qquad = \lambda_1 e^{-z_1} + \lambda_2 e^{-z_2}.
\end{aligned} \tag{A.40}$$

Let E_j, $j = 1, \cdots, N$ be the events whose probabilities are being summed in (A.40). The assertion (A.38) corresponds to a limiting statement about $P(E_1 \cdots E_N)$. If the E_j were independent assertion (A.38) would follow readily from (A.40). Let $\tilde{P}$ be the measure which makes the vectors $(Y_T(a_1), Y_T(a_1 + ax^{-2/\alpha}), \cdots, Y_T(a_1 + ax^{-2/\alpha}[tx^{2/\alpha}/a]))$, $(Y_T(a_2), \cdots, Y_T(a_2 + ax^{-2/\alpha}[tx^{2/\alpha}/a]))$, $\cdots$, $(Y_T(a_N), \cdots, Y_T(a_N + ax^{-2/\alpha}[tx^{2/\alpha}/a])$ independent and otherwise agrees with P.

To conclude the proof of the theorem we need to show that.

$$\lim_{\varepsilon\to 0} \overline{\lim}_T |(P - \tilde{P})(E_1 \cdots E_N)| = 0 . \tag{A.41}$$

To do this apply the following modification of Lemma 4.1 of [1].

LEMMA A4. *Let*

$$\phi(x, y, p) = \frac{1}{2\pi(1 - p^2)^{\frac{1}{2}}} \exp - \frac{(x^2 - 2pxy + y^2)}{2(1 - p^2)} . \tag{A.42}$$

Let $\Sigma_1 = |r_{ij}|$, $\Sigma_2 = |s_{ij}|$ *be* $k \times k$ *nonnegative semi-definite matrices with* $r_{ii} = s_{ii} = 1$ *for all* i. *Let* $\mathbf{X} = (X_1, \cdots, X_k)$ *be a mean* 0 *Gaussian vector with covariance matrix* Σ_1 *or* Σ_2. *Let* $u_1, \cdots, u_k$ *be nonnegative numbers and* $u = \min_j u_j$. *Then,*

$$\begin{aligned} &|P_{\Sigma_1}[X_j \leqq u_j, 1 \leqq j \leqq k] - P_{\Sigma_2}[X_j \leqq u_j, 1 \leqq j \leqq k]| \\ &\qquad \leqq 4 \textstyle\sum_{i,j} |\int_{s_{ij}}^{r_{ij}} \phi(u, u; \lambda)\, d\lambda| . \end{aligned} \tag{A.43}$$

PROOF. By the usual argument (see [1] page 931) the left-hand side of (A.43) is bounded by, $4 \sum_{i,j} |\int_{s_{ij}}^{r_{ij}} \phi(u_i, u_j; \lambda)\, d\lambda|$. But, by an elementary inequality

$$x^2 - 2pxy + y^2 \geqq \frac{(1 - p)}{2} (x + y)^2 . \tag{A.44}$$

Thus,

$$\phi(u_i, u_j, \lambda) \leqq \phi\left(\frac{u_i + u_j}{2}, \frac{u_i + u_j}{2}, \lambda\right) \leqq \phi(u, u, \lambda) . \tag{A.45}$$ □

Take $X_1 = Y_T(a_1) - \mu_T(a_1)$, $X_2 = -Y_T(a_1) + \mu_T(a_1)$ etc., $k = 2N[tx^{2/\alpha}/a]$, $|r_{ij}|$ corresponding to the distribution of $\mathbf{X}$ under P, $|s_{ij}|$ corresponding to $\tilde{P}$, $u_1 = x - \mu_T(a_1)$, $u_2 = y + \mu_T(a_1)$ etc. Evidently,

$$u = (2 \log T)^{\frac{1}{2}} + O((\log T)^{-\frac{1}{2}}) . \tag{A.46}$$

It is clear now that we can apply to the bound of (A.43) exactly the same analysis as that given by Berman on pages 933–936 of [1] to arrive at the conclusion of the theorem.

Note. By applying the more refined analysis of Pickands [11] pages 64–72 we can show that the conclusion of the theorem also holds if (vi) is replaced by,

$$\lim_{t\to\infty} r(t) \log t = 0 . \tag{A.47}$$

Unfortunately, the analysis of Berman appears to only yield the conclusion under the stronger

$$r(t)[\log t]^{2/\alpha} \to 0 . \tag{A.48}$$

We do not enter into this further since (vi) is what we need for Theorems 1.1 and 1.2.

5. Appendix B. Miscellanea.

THEOREM B1. *Let* w *be an absolutely continuous square integrable function with* *a square integrable derivative* w'. *Let*,

$$r(t) = \int w(t+s)w(s)\,ds\,. \tag{B.1}$$

Then r *is twice differentiable and*

$$r''(t) = -\int w'(t+s)w'(s)\,ds\,. \tag{B.2}$$

PROOF. We first show that

$$r'(t) = \int w'(t+s)w(s)\,ds = \int w(s-t)w'(s)\,ds\,. \tag{B.3}$$

Let $\hat{w}$, $\hat{w}'$ be the Fourier transforms of w, w'. Then by Parseval,

$$\frac{r(t+h)-r(t)}{h} = \frac{1}{2\pi}\int \frac{(e^{-i(t+h)u} - e^{-itu})}{h}|\hat{w}(u)|^2\,du\,. \tag{B.4}$$

Applying the dominated convergence theorem we obtain the existence of r' given by

$$r'(t) = -\frac{i}{2\pi}\int e^{-itu}u|\hat{w}(u)|^2\,du = \int w'(t+s)w(s)\,ds\,.$$

Similarly

$$\begin{aligned}\frac{r'(t+h)-r'(t)}{h} &= \int \frac{w(s-t-h)-w(s-t)}{h}w'(s)\,ds\\ &= \frac{i}{2\pi}\int\left(\frac{e^{i(t+h)u} - e^{itu}}{h}\right)u|\hat{w}(u)|^2\,du\\ &\to -\frac{1}{2\pi}\int e^{itu}u^2|\hat{w}(u)|^2\,du = -\int w'(s-t)w'(s)\,ds\,.\end{aligned} \tag{B.5}$$

The theorem follows. Note that $r'(0) = 0$ from (B4) since $|\hat{w}|$ is symmetric.

THEOREM B2. *Let* w *be absolutely continuous on* $[-A, A]$ *and* 0 *otherwise. Then* r *has left and right derivatives at* 0 *and*

$$r_+'(0) = -r_-'(0) = -\tfrac{1}{2}(w^2(A) + w^2(-A))\,. \tag{B.6}$$

PROOF. Write, for $h > 0$,

$$\begin{aligned}&\int \frac{w(s+h)-w(s)}{h}w(s)\,ds\\ &\quad= \int_{-A}^{A-h}\left[\frac{1}{h}\int_s^{s+h} w'(z)\,dz\right]w(s)\,ds - \frac{1}{h}\int_{A-h}^{A} w^2(s)\,ds\\ &\quad\to \int_{-A}^{A} w'(s)w(s)\,ds - w^2(A) = -\tfrac{1}{2}(w^2(A) + w^2(-A))\end{aligned} \tag{B.7}$$

by arguing as in Theorem A1 and using Lebesgue's theorem. Since $r(-t) = r(t)$ the result follows.

THEOREM B3. (Pickands) *If* $H_\alpha(n, a)$, $\bar{H}_\alpha(n)$ *are defined as in* (A.17), (A.19) *then* (A.20) *holds.*

PROOF. Suppose first that $0 < \alpha < 2$. Let for $\gamma > 0$,

$$\bar{H}_\alpha(n,\gamma) = \int_{-\infty}^{\infty} e^s[\max_{0\le t\le n} Y(t) > s+\gamma]\,ds = e^{-\gamma}\bar{H}_\alpha(n)\,. \tag{B8}$$

Then

$$\begin{aligned}
&\frac{1}{n}|H_\alpha(n,a)-\bar{H}_\alpha(na,\gamma)| \\
&\quad\le \frac{1}{n}\Big[\textstyle\int_{-\infty}^{\infty} e^s P[\max_{0\le t\le na} Y(t) > s+\gamma,\ \max_{0\le k\le n} Y(ka)\le s]\,ds \\
&\qquad + \textstyle\int_{-\infty}^{\infty} e^s P[s < \max_{0\le t\le na} Y(t) \le s+\gamma]\,ds\Big] \qquad \text{(B.9)}\\
&\quad\le \frac{1}{n}\textstyle\sum_{k=0}^{n-1}\int_{-\infty}^{\infty} e^s P[Y(ka)\le s,\ \max_{ka\le t\le (k+1)a} Y(t) > s+\gamma]\,ds \\
&\qquad + \frac{1}{n}[\bar{H}_\alpha(na)-\bar{H}_\alpha(na,\gamma)]\,.
\end{aligned}$$

If the summands on the right of the first term of (B.9) are denoted by $A(k,\gamma,a)$ then,

$$\text{(B.10)}\qquad \begin{aligned} A(k,\gamma,a) &= \textstyle\int_{-\infty}^{\infty} e^s \int_{-\infty}^{s} \tau(z,ka) \\ &\quad \times P[\max_{0\le t\le a} Y(t+ka) > s+\gamma \mid Y(ka)=z]\,dz\,ds \end{aligned}$$

where $\tau(z,ka)$ is the density of $Y(ka)$. After some manipulation we obtain

$$\text{(B.11)}\qquad \begin{aligned} A(k,\gamma,a) &= \textstyle\int_{-\infty}^{\infty}\phi(w)\int_0^{\infty} e^s P[\max_{0\le t\le a}(Y(t+ka)-Y(ka)) > s+\gamma \mid Y(ka) \\ &= w+(ka)^\alpha]\,ds\,dw\,. \end{aligned}$$

As $k\to\infty$, the finite dimensional conditional distributions of $Y(t+ka)-Y(ka)$ given $Y(ka)=w+(ka)^\alpha$ tend for each w to those of $Y(t)$, $0\le t\le a$. Arguing as in Lemma A1 we conclude that,

$$\text{(B.12)}\qquad \lim_k A(k,\gamma,\alpha) = A(\gamma,\alpha) = \textstyle\int_0^{\infty} e^s P[\max_{0\le t\le a} Y(t) > s+\gamma]\,ds\,.$$

Let $Y^*(t) = Y(t)+|t|^\alpha$. Then,

$$\text{(B.13)}\qquad \begin{aligned} A(\gamma,\alpha) &\le \textstyle\int_0^{\infty} e^s P[\max_{0\le t\le a} Y^*(t) > s+\gamma]\,ds \\ &= \textstyle\int_0^{\infty} e^s P[\max_{0\le t\le 1} Y^*(t) > (s+\gamma)a^{-\alpha/2}]\,ds \\ &= a^{\alpha/2}e^{-\gamma}\textstyle\int_{\gamma a^{-\alpha/2}}^{\infty} e^{wa^{\alpha/2}} P[\max_{0\le t\le 1} Y^*(t) > w]\,dw\,. \end{aligned}$$

Applying Fernique's estimate the right-hand side of (B.13) is $O(\exp -a^{-\alpha/2})$ for every $\gamma>0$. We conclude that,

$$\text{(B.14)}\qquad \begin{aligned} &\limsup_a \limsup_n \frac{1}{na}|H_\alpha(n,a)-\bar{H}_\alpha(na,\gamma)| \\ &\qquad\qquad \le (1-e^{-\gamma})\limsup_a \limsup_n \frac{\bar{H}_\alpha(na)}{na} \end{aligned}$$

for every $\gamma>0$. Since,

$$\text{(B.15)}\qquad \begin{aligned} &P[\max_{0\le t\le n} Y(t) > s] \\ &\quad\le \textstyle\sum_{k=0}^{n-1} P[\max_{k\le t\le k+1} Y(t) > s] \\ &\quad\le \textstyle\sum_{k=0}^{n-1}\{P[Y(k)\le s,\ \max_{k\le t\le k+1} Y(t) > s] + P[Y(k) > s]\}\,, \end{aligned}$$

it is easy to see that,

$$\text{(B.16)} \qquad \sup_{x \geq 1} \frac{\bar{H}_\alpha(x)}{x} < \infty .$$

Hence,

$$\text{(B.17)} \qquad \lim_a \lim\sup_n \frac{1}{na} |H_\alpha(n, a) - \bar{H}_\alpha(na)| = 0 .$$

But from the argument of Lemma A3 it is clear that for every $a > 0$,

$$\text{(B.18)} \qquad \lim\sup_n \frac{H_\alpha(n, a)}{na} \leqq \lim\inf_n \frac{\bar{H}_\alpha(na)}{na} .$$

The theorem follows for $0 < \alpha < 2$. For $\alpha = 2$ we can use the representation $Y(t) = 2^{\frac{1}{2}}tZ - t^2$ where Z is a standard normal deviate. Evidently,

$$\text{(B.19)} \qquad \begin{aligned} \max_{0 \leqq s \leqq na/2^{\frac{1}{2}}} Y(s) &= \frac{Z^2}{2} \quad \text{if} \quad 0 \leqq Z < \frac{na}{2^{\frac{1}{2}}} \\ &= naZ - \frac{n^2a^2}{2} \quad \text{otherwise} . \end{aligned}$$

It follows that,

$$\text{(B.20)} \qquad \begin{aligned} &\frac{1}{na}\left|\bar{H}_2\left(\frac{na}{2^{\frac{1}{2}}}\right) - H_2\left(n, \frac{a}{2^{\frac{1}{2}}}\right)\right| \\ &\quad \leqq \frac{1}{na} \int_0^{n^2a^2} e^{s/2} P[s^{\frac{1}{2}} < Z < (s + a^2)^{\frac{1}{2}}]\, ds \sim 2(1 - e^{-a^2/2}) \end{aligned}$$

by standard arguments. The theorem now follows generally.

REFERENCES

[1] BERMAN, S. N. (1971). Asymptotic independence of the numbers of high and low level crossings of stationary Gaussian processes. *Ann. Math. Statist.* **42** 927–946.

[2] BERMAN, S. N. (1971). Maxima and high level excursions of stationary Gaussian processes. *Trans. Amer. Math. Soc.* To appear.

[3] BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

[4] BREIMAN, L. (1969). *Probability.* Addison-Wesley, Reading.

[5] BRILLINGER, D. (1969). An asymptotic representation of the sample distribution function. *Bull. Amer. Math. Soc.* **75** 545–547.

[6] DOOB, J. L. (1953). *Stochastic Processes.* Wiley, New York.

[7] EPANECHNIKOV, V. A. (1969). Nonparametric estimates of a multivariate probability density. *Theor. Probability Appl.* **14** 153–158.

[8] FERNIQUE, X. (1970). Intégrabilité des vecteurs gaussiens. *C.R. Acad. Sci. Paris* **270** A 1698–99.

[9] HAJEK, J. and SIDAK, Z. (1967). *Theory of Rank Tests.* Academic Press, New York.

[10] MANN, H. B. and WALD, A. (1942). On the choice of the number of class intervals in the application of the χ^2 test. *Ann. Math. Statist.* **13** 306–317.

[11] PICKANDS, J. S. III (1960). Upcrossing probabilities for Gaussian processes. *Trans. Amer. Math. Soc.* **145** 51–73.

[12] QUALLS, C. and WATANABE, H. (1971). Asymptotic properties of Gaussian processes. Tech. Report No. 736, Univ. of North Carolina, Chapel Hill.

[13] ROSENBLATT M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

[14] STECK, G. P. (1957). Limit theorems for conditional distributions. *Univ. California Publ. Statist.* **2** 237–284.

[15] WATSON, G. S. (1967). Some problems in the statistics of directions. *Bull. I.S.I.* **17** 374–385.

[16] WOODROOFE, M. (1967). On the maximum deviation of the sample density. *Ann. Math. Statist.* **38** 475–481.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
LA JOLLA, CALIFORNIA 92037

Sankhyā : The Indian Journal of Statistics
1988, Volume **50**, Series A, Pt. 3, pp. 381-393.

ESTIMATING INTEGRATED SQUARED DENSITY DERIVATIVES : SHARP BEST ORDER OF CONVERGENCE ESTIMATES*

By P. J. BICKEL
University of California at Berkeley
and
Y. RITOV
The Hebrew University of Jerusalem

SUMMARY. Estimation of the integral of the square of a derivative of the probability density function is considered. The estimators we propose and their properties are a function of the amount of smoothness assumed. The rate of convergence of the appropriate estimator is shown to be optimal given the amount of smoothness assumed. In particular the appropriate estimator achieves the information bound when estimation at an $n^{-1/2}$ rate is possible.

1. Introduction

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d., each with distribution function F. Let $f(.)$ be the probability density function of F, $f^{(k)}$ its k-th derivative and $\theta_k(F) = \int \{f^{(k)}(x)\}^2 \, dx$. These functionals appear in the asymptotic variance of the Wilcoxon statistic and in the asymptotics of the integrated M.S.E. for kernel density estimates. Discussion of the estimation of θ_k and similar parameters appear in Schweder (1975), Hasminskii and Ibragimov (1978), Pfanzagl (1982), Prakasa Rao (1983), Donoho and Liu (1987) and Hall and Marron (1987).

Ritov and Bickel (1987) show that the standard semiparametric information bound for the estimation of $\theta_0(F)$ fails to give an achievable rate of convergence. In fact, the information is strictly positive when f is bounded, promising that the $n^{-1/2}$ rate is achievable. Nevertheless, there is no rate that can be achieved uniformly in small compact neighborhoods (in the total variation norm) of a given distribution. Moreover, even if the uniformity requirement is dropped then for any sequence of estimates $\{\hat{\theta}_k\}$ there exists an (unknown) point F such that $n^{\gamma}(\hat{\theta}_k - \theta_k(F))$ doesn't converge to 0 for any $\gamma > 0$.

In this paper we consider classes of F which satisfy Hölder conditions on $f^{(m)}$ for suitable m. We establish the rate achievable under these condi-

*Research supported by Office of Naval Research N00014-80-C-0163.

AMS (1980) *subject classification :* 62G05, 62G20,

Key words and phrases : estimation, density derivatives, semiparametric information bound, sharp rate convergence.

*An invited paper to Commemorate the 50-th volume of Sankhyā.

tions and exhibit estimators that achieve these rates. Our estimators converge uniformly and when improvement is possible faster than similar estimators suggested by Schweder (1975), Hasminskii and Ibragimov (1978), and Hall and Marron (1987). In particular we need to assume weaker Hölder conditions to obtain $n^{-1/2}$ rates and efficient estimators.

We believe that our proof of the best achievable rates is novel in that it cannot be reduced to considering a sequence of simple vs. simple testing problems and in effect requires the use of composite hypotheses of growing size. Note that θ_k can be estimated at the $n^{-1/2}$ rate in any fixed regular finite dimensional submodel.

2. Main results : the estimators and their properties

Let $\theta_k(F) = \int \{f^{(k)}(x)\}^2\, dx$ where f is the (continuous) density of the distribution F. (In general we denote distribution functions by F or F_n and their densities by f or f_n respectively.) Let $\alpha > 0$, m be a nonnegative integer and $g(\cdot) \in L_2 \bigcap L_\infty$. Suppose $X_1, \ldots, X_n$ is a random sample from F. How well can $\theta_k(F)$ be estimated if it is known a priori only that $F \in \boldsymbol{F}_{m,\alpha,g}$ where $\boldsymbol{F}_{m,\alpha,g} = \{F : |f^{(m)}(x) - f^{(m)}(x+\xi)| \leqslant g(x)|\xi|^\alpha$ for all x real $|\xi| < 1\}$?

We begin by suggesting a family of estimators. Let $h_\sigma(x) = \sigma^{-1} h(x/\sigma)$ where h is a kernel with the following properties :

$$\begin{aligned}
&h \text{ is symmetric about zero,}\\
&h(x) = 0 \text{ for } |x| > 1,\\
&\int h(x)dx = 1,\\
&\int x^i h(x)dx = 0, \quad i = 1, 2, \ldots, \max\{k, m-k\}
\end{aligned}$$

and h has $2k+1$ derivatives.

Divide the sample into two subsamples $X_1, \ldots, X_{n_1}$ and $X_{n_1+1}, \ldots, X_n$ with comparable sizes (i.e. n_1/n is bounded away from 0 and 1). Let $\hat{F}_1$ and $\hat{F}_2$ be the empirical distribution functions of each subsample respectively. Define, $\hat{f}_i(x) = \int h_\sigma(x-y)d\hat{F}_i(y)$, $i = 1, 2$. The dependence of $\hat{f}_i$ on σ is left implicit. Consider the following estimator of θ_0.

$$\hat{\theta}_0^*(X_1, \ldots, X_n\,; \sigma) = \frac{n_1}{n}\,\hat{\theta}_{01}^* + \frac{n_2}{n}\,\hat{\theta}_{02}^* \qquad \ldots \quad (2.1)$$

where $n = n_1 + n_2$

$$\begin{aligned}
&\hat{\theta}_{01}^*(X_1, \ldots, X_n; \sigma)\\
&= \int \hat{f}_2^2(x)dx + 2n_1^{-1} \sum_{i=1}^{n_1} (\hat{f}_2(X_i) - \int \hat{f}_2^2(x)dx) + \frac{1}{n_2} \int h_\sigma^2(x)dx\\
&= 2\int h_\sigma(x-t)d\hat{F}_1(t)d\hat{F}_2(x) - n_2^{-2} \sum_{n_1+1 \leqslant i \neq j \leqslant n}{}' \int h_\sigma(x-X_i)h_\sigma(x-X_j)dx
\end{aligned}$$

$$\ldots \quad (2.2)$$

and $\hat{\theta}^*_{02}$ is obtained by interchanging the roles of the two subsamples in $\hat{\theta}^*_{01}$. The first two terms of $\hat{\theta}^*_{01}$ can be recgonized as Hasminskii and Ibragimov's estimate of this parameter which they show is efficient in $F_{0,\alpha,M}$ if $\alpha > 1/2$. This is the, by now, familiar one step estimate (see Bickel, 1982 ; Schick, 1986) using the estimated influence function $2(\hat{f}_2 - \int \hat{f}_2^2(x)dx)$. The last term in (2.2) removes the pure known bias component, $n_2^{-2} \sum_{i=n_1+1}^{n} \int h_\sigma^2(x-X_i)dx$ from

$$\int \hat{f}_2^2(x)dx = n_2^{-2} \sum_{i,j} \int h_\sigma(x-X_i)h_\sigma(x-X_j)dx. \qquad \dots \quad (2.3)$$

Curiously enough this simple debiasing leads to efficient estimation in $\boldsymbol{F}_{0,\alpha,M}$ for $\alpha > 1/4$ and (uniformly) $\sqrt{n}$ consistent estimation on $\boldsymbol{F}_{0,1/4,M}$. Moreover, $\sqrt{n}$ consistent estimation is shown to be impossible for $\alpha < 1/4$. More generally, if f has $2k$ continuous derivatives,

$$\begin{aligned} \theta_k(F) &= (-1)^k \int f^{(2k)}(x)f(x)dx \\ &= (-1)^k E_F(f^{(2k)}(X)). \end{aligned}$$

This suggests, by the same process as above, estimates $\hat{\theta}^*_{k1}$, $\hat{\theta}^*_{k2}$ and $\hat{\theta}^*_k$. For convenience we replace $\hat{\theta}^*_{01}$ by $\hat{\theta}_{01}$ where n_2^{-2} in (2.2) is replaced by $[n_2(n_2-1)]^{-1}$ and similar replacements are made in $\hat{\theta}^*_{02}$ and more generally $\hat{\theta}^*_k$. So the estimate we study is

$$\begin{aligned} \hat{\theta}_k(X_1, \dots, X_n; \sigma) &= 2(-1)^k \int h_\sigma^{(2k)}(x-t)d\hat{F}_1(t)d\hat{F}_2(x) \\ &- n_2[n\, n_1(n_1-1)]^{-1} \sum_{1 \leqslant i<j \leqslant n_1} \int h_\sigma^{(k)}(x)-X_i)h_\sigma^{(k)}(x-X_j)dx \\ &- n_1[n\, n_2(n_2-1)]^{-1} \sum_{n_1+1 \leqslant i<j \leqslant n} \int h_\sigma^{(k)}(x-X_i)h_\sigma^{(k)}(x-X_j)dx. \qquad \dots \quad (2.4) \end{aligned}$$

Our main results are summarized in the following two theorems. In the first we describe the performance of $\hat{\theta}_k$ in terms of the assumed family $\boldsymbol{F}_{m,\alpha,g}$. The rate of convergence of $\hat{\theta}_k$ to $\theta_k(F)$ is a function of $m+\alpha$ and $\hat{\theta}_k$ is "efficient" when $m+\alpha > 2k+1/4$. In the second theorem we show that the rates given in the first theorem are, essentially, the best possible.

Theorem 1 : *Let* $\{F_1, F_2, \dots\} \subset \boldsymbol{F}_{m,\alpha,g}$ *where* $0 \leqslant \alpha < 1$, $m+\alpha > k$ *and* $g \in L_2 \cap L_\infty$. *Let* $X_{n1}, \dots, X_{nn}$ *be i.i.d.*, $X_{n1} \sim F_n$ *and let* $\hat{\theta}_k = \hat{\theta}_k(X_{n1}, \dots, X_{nn}; \sigma_n)$ *where* $\sigma_n = n^{-2/(1+4m+4\alpha)}$.

(i) *If* $m+\alpha > 2k+1/4$ *then*

$$\sqrt{n}\Big[\, \hat{\theta}_k - \theta_k(F_n) - \frac{2}{n} \sum_{i=1}^{n} \{(-1)^k f_n^{(2k)}(X_{ni}) - \theta_k(F_n)\}\Big] \longrightarrow 0. \qquad \dots \quad (2.5)$$

Let $I_k(F_n) = [Var\{f_n^{(2k)}(X_{n1})\}]^{-1}$. *Then,* $n\, I_k(F_n) E\, \{\hat{\theta}_k - \theta_k(F_n)\}^2 \to 1$ *and* $L\{\sqrt{n}\, I_k^{1/2}(F_n)\, (\hat{\theta}_k - \theta_k(F_n))\} \to N(0, 1)$ *provided* $\limsup_n I_k(F_n) < \infty$.

(ii) *If* $k < m+\alpha \leqslant 2k+1/4$ *then* $n^{2\gamma}E\{\hat{\theta}_k-\theta_k(F_n)\}^2$ *is bounded when* $\gamma = 4(m+\alpha-k)/(1+4m+4\alpha)$.

We conjecture, but have not checked the details, that it is possible to estimate σ by cross validation to obtain an estimate $\hat{\theta}_k^* = \hat{\theta}_k(X_{n1}, \ldots, X_{nn} ; \hat{\sigma}_n)$ which does not depend on m and α but is equivalent to $\hat{\theta}_k$ which does so depend through σ_n given in the statement of Theorem 1.

Theorem 2 : (i) *The information bound (in the sense of Khoshevnik and Levit (1976)) for non parametric estimation of* $\theta_k(F)$, $F \in \boldsymbol{F}_{2k,\alpha,g}$ *is given by* $I_k(F)$ *as defined in Theorem* 1.

(ii) *Suppose* $k < m+\alpha \leqslant 2k+1/4$. *Then there is a small compact set* $\boldsymbol{F}^* \subsetneq \boldsymbol{F}_{m,\alpha,g}$ *such that for any* $c_n \to \infty$ *and any sequence of estimators* $T_1, T_2, \ldots, T_n = T_n(X_1, \ldots, X_n)$, $X_1, X_2, \ldots, X_n$ *iid*, $X_1 \sim F$:

$$\liminf_{n} \sup_{F \in \boldsymbol{F}^*} P_F\{c_n n^{\gamma} | T_n-\theta_k(F) | \geqslant 1\} = 1 \qquad \ldots \quad (2.6)$$

where $\gamma = 4(m+\alpha-k)/(1+4m+4\alpha)$. *Moreover* $\boldsymbol{F}^*$ *can be constructed so that its only accumulation point is any specified* $F_0 \in \boldsymbol{F}_{m,\alpha,g}$.

The proof of the first part of Theorem 2 is quite standard and follows essentially the discussion in Hasminskii and Ibragimov (1978). The proof of the second part of the Theorem is an extension of the ideas presented in Ritov and Bickel (1987). In our problem, θ_0 can be estimated at the $n^{-1/2}$ rate in any one dimensional sub model of $\boldsymbol{F}_{m,\alpha,g}$ and the information bound of Theorem 2i) is the best bound that can be achieved using these techniques. Yet for $m+\alpha < 2k+1/4$ this bound is unachievable by uniformly $n^{1/2}$ consistent estimates. In fact, for $m+\alpha < 2k+1/4$ no uniformly $n^{1/2}$ consistent estimate exists. Even uniformity can be dropped—see Ritov and Bickel (1987), Theorem 1. Our proof is based on the demonstration of a sequence of difficult multiparameter Bayesian problems.

3. Proofs

We begin the proofs with the following technical lemma whose own proof is postponed to the end of the section.

Lemma 1 : *Let* α, m *and* g *be such that* $\alpha > 0$ $m \geqslant 0$ *and* $g \in L_\infty$. *Then* $sup\{|f^{(i)}(x)| : x, F \in \boldsymbol{F}_{m,\alpha,g}\} < \infty$, $i = 0, 1, \ldots, m$.

Proof of Theorem 1 : Evidently to establish Theorem 1 it is enough to consider the asymmetric estimate

$$\hat{\theta}_{k2} = 2(-1)^k \int\int h_\sigma^{(2k)}(x-t)\, d\hat{F}_1(t)\, d\hat{F}_2(x).$$

$$-\ 2\{n_1(n_1-1)\}^{-1} \sum_{1\leqslant i<j\leqslant n_1} \int h_\sigma^{(k)}(x-X_{ni})\, h_\sigma^{(k)}(x-X_{nj})\, dx.$$

We begin by estimating the conditional bias

$$E(\hat{\theta}_{k2}\,|\,\hat{F}_1)-\theta_k(F_n) = 2(-1)^k \int \hat{f}_1^{(2k)}(x)\, f_n(x)\, dx$$

$$-2\,\{n_1(n_1-1)\}^{-1} \sum_{i=1}^{n_1}\sum_{j=1}^{i-1} \int h_\sigma^{(k)}(x-X_{ni})\, h_\sigma^{(k)}(x-X_{nj})\, dx - \int \{f_n^{(k)}(x)\}^2\, dx.$$

But

$$(-1)^k \int \hat{f}_1^{(2k)}(x) f_n(x) dx = \int \hat{f}_1^{(k)}(x) f_n^{(k)}(x) dx$$

$$= n_1^{-1} \sum_{i=1}^{n_1} \int h_\sigma^{(k)}(x-X_{ni}) f_n^{(k)}(x) dx$$

$$= \{n_1(n_1-1)\}^{-1} \sum_{i=1}^{n_1} \sum_{1\leqslant j\neq i\leqslant n_1} \int h_\sigma^{(k)}(x-X_{ni}) f_n^{(k)}(x) dx.$$

Hence

$$E(\hat{\theta}_{k2}\,|\,\hat{F}_1)-\theta_k(F_n) = -2\{n_1(n_1-1)\}^{-1} \sum_{i=1}^{n_1}\sum_{j=1}^{i-1} \int\{h_\sigma^{(k)}(x-X_{ni})-f_n^{(k)}(x)\}$$

$$\{h_\sigma^{(k)}(x-X_{nj})-f_n^{(k)}(x)\}dx. \quad \ldots \quad (3.1)$$

We obtain from (3.1) that

$$E\,\hat{\theta}_{k2}-\theta_k(F_n) = \int \{f_{n\sigma}^{(k)}(x)-f_n^{(k)}(x)\}^2 dx \quad \ldots \quad (3.2)$$

where $f_{n\sigma} = f_n * h_\sigma$.

But

$$f_{n\sigma}^{(k)}(x)-f_n^{(k)}(x) = \int h(t)\,\{f_n^{(k)}(x+\sigma t)-f_n^{(k)}(x)\}dt$$

$$= \int h(t)\left\{\sum_{i=1}^{m-k-1} \frac{f_n^{(k+i)}(x)}{i\,!}\,\sigma^i t^i\right\}dt \quad \ldots \quad (3.3)$$

$$+ \int h(t)\, \frac{1}{(m-k)!}\, \{f_n^{(m)}(x+\sigma^* t)-f_n^{(m)}(x)\}\sigma^{m-k} t^{m-k} dt,$$

where $0 \leqslant \sigma^* \leqslant \sigma$. The first term in the RHS of (3.3) is null by the construction of h. Since $F_n \in \boldsymbol{F}_{m,a,g}$ we can bound the integrand in the second term and obtain :

$$|f_{n\sigma}^{(k)}(x)-f_n^{(k)}(x)| \leqslant g(x)\sigma^{m+a-k} \int |t|^{m+a-k}\, |h(t)|\, dt. \quad \ldots \quad (3.4)$$

A 3–13

Combine (3.2) and (3.4) to conclude that

$$|E\,\hat{\theta}_{k2}-\theta_k(F_n)| \leqslant \|g\|_2^2\, n^{-4(m+\alpha-k)/(1+4m+4\alpha)} \left(\int |t|^{m+\alpha-k}\,|h(t)|\,dt\right)^2. \quad \ldots \quad (3.5)$$

Next we estimate var $(E(\hat{\theta}_{k2}\,|\,\hat{F}_1))$. Note that $E(\hat{\theta}_{k2}\,|\,\hat{F}_1)$ was written in (3.1) as a U-statistic, $E(\hat{\theta}_{k2}\,|\,F_1)-\theta_k(F_n) = 2\{n_1(n_1-1)\}^{-1}\sum_{i=1}^{n_1}\sum_{j=1}^{i-1} U(X_{ni}, X_{nj}')$ say.

By standard U-statistic theory,

$$\text{var}\,\{E(\hat{\theta}_{k2}\,|\,\hat{F}_1)\} = n^{-1}\{O(\text{var}[E\,(U\,(X_{n1}, X_{n2})\,|\,X_{n1}])$$
$$+O\,(n^{-1}\ \text{var}\ \ U(X_{n1}, X_{n2}))\}. \quad \ldots \quad (3.6)$$

Now

$$E\,U\,(x, X_{n2}) = \int \{h_\sigma^{(k)}(t-x)-f_n^{(k)}(t)\}\,\{f_{n\sigma}^{(k)}(t)-f_n^{(k)}(t)\}\,dt$$
$$= \int \delta(t)\,\{h_\sigma^{(k)}(t-x)-f_n^{(k)}(t)\}dt,$$

say. Hence,

$$\text{var}\ [E\{U(X_{n1}, X_{n2})\,|\,X_{n1}\}] = E[\int\delta(x)\,\{h_\sigma^{(k)}(x-X_{n1})-f_{n\sigma}^{(k)}(x)\}\,dx]^2$$
$$= E\int\int \delta(y)\,\delta(x)\,\{h_\sigma^{(k)}(y-X_{n1})-f_{n\sigma}^{(k)}(y)\}\,\{h_\sigma^{(k)}(x-X_{n1})-f_{n\sigma}^{(k)}(x)\}\,dx\,dy$$
$$\leqslant \int\int \delta(y)\,\delta(x)\,\int h_\sigma^{(k)}(y-t)h_\sigma^{(k)}(x-t)f_n\,(t)\,dt\,dx\,dy$$
$$= \int \{\int\delta(x)h_\sigma^{(k)}(x-t)dx\}^2 f_n\,(t)\,dt$$
$$\leqslant \|\delta\|_\infty^2\,\sigma^{-2k}\,\{\int |\,h^{(k)}\,(x)\,|\,dx\}^2 = O(\sigma^{2(m+\alpha-2k)}) \quad \ldots \quad (3.7)$$

by (3.4). At the same time, the random variable $\int h_\sigma^{(k)}(x-X_{n1})h_\sigma^{(k)}(x-X_{n2})dx$ is bounded by $\sigma^{-2k-1}\,\|h^{(k)}\|_2^2$ and is equal to zero unless $|X_{n1}-X_{n2}| \leqslant 2\sigma$. Since f_n is bounded this last event has probability of the same order as σ.

Hence

$$\text{var}\quad \{\int h_\sigma^{(k)}(x-X_{n1})h_\sigma^{(k)}(x-X_{n2})dx\} = O\,(\sigma.\,\sigma^{-4k-2}).$$

Since $|\int f_n^{(k)}(x)\,.\,h_\sigma^{(k)}(x-X_{n1})dx| \leqslant \|f_n^{(k)}\|_\infty\,\sigma^{-k}\int\,|h^{(k)}(x)|\ dx$ we conclude that var $\{U\,(X_{n1}, X_{n2})\}$

$$= \text{var}[\int \{h_\sigma^{(k)}(x-X_{n1})h_\sigma^{(k)}(x-X_{n2})-f_n^{(k)}(x)h_\sigma^{(k)}(x-X_{n1})-f_n^{(k)}(x)h_\sigma^{(k)}(x-X_{n2})\}\,dx)]$$
$$= O(\sigma^{-4k-1}). \quad \ldots \quad (3.8)$$

We obtain from (3.1), (3.4), (3.6), (3.7), and (3.8) that

$$\text{var}\,\{E\,(\hat{\theta}_{k2}\,|\,\hat{F}_1)\} = O(n^{-1}\sigma^{2(m+\alpha-2k)}+n^{-2}\sigma^{-4k-1})$$
$$= O(n^{-8(m+\alpha-k)/(1+4m+4\alpha)})$$

for σ given in the statement of Theorem 1. Hence (3.5) implies that

$$E\{E(\hat{\theta}_{k2} \mid \hat{F}_1) - \theta_k(F_n)\}^2 = O(n^{-8(m+\alpha-k)/(1+4m+4\alpha)}). \qquad \dots \quad (3.9)$$

We have proved that $E(\hat{\theta}_{k2} \mid \hat{F}_1) - \theta_k(F_n)$ is of the right order (in particular it is $o_p(n^{-1/2})$ if $m+\alpha > 2k+1/4$). We turn to the investigation of the behaviour of $\hat{\theta}_{k2} - E(\hat{\theta}_{k2} \mid \hat{F}_1)$. This will be carried on separately for the two cases : $2k+1/4 < m+\alpha$ and $k < m+\alpha \leqslant 2k+1/4$.

(i) Suppose $2k+1/4 < m+\alpha$. In the light of (3.9) we need only to consider the conditional variance of $\hat{\theta}_{k2}$ given the first sub sample. But, given $X_{n1}, \dots, X_{nn_1}$, $\hat{\theta}_{k2}$ is just a sum of *i.i.d.* random variables, hence

$$\operatorname{var}\left\{\hat{\theta}_{k2} - \frac{2(-1)^k}{n-n_1} \sum_{i=n_1+1}^{n} f_n^{(2k)}(X_{ni}) + \theta_k(F_n) \mid \hat{F}_1\right\}$$

$$\leqslant \frac{4}{n-n_1} \int \{\hat{f}_1^{(2k)}(x) - f_n^{(2k)}(x)\}^2 f_n(x)\, dx.$$

So
$$E \operatorname{var}\{\hat{\theta}_{k2} - 2(-1)^k \textstyle\int f_n^{(2k)}(x)\, d\hat{F}_2(x) + \theta_k(F_n) \mid \hat{F}_1\}$$

$$\leqslant \frac{4}{n-n_1} \int \{f_{n\sigma}^{(2k)}(x) - f_n^{(2k)}(x)\}^2 f_n(x) dx + \frac{4}{n-n_1} \int \{\operatorname{var} \hat{f}_1^{(2k)}(x)\} dx.$$

$$= o_P(n^{-1}). \qquad \dots \quad (3.10)$$

Now (3.9) and (3.10) imply the validity of (2.5). Since by Lemma 1, f_n is uniformly bounded, the first part of Theorem 1 follows.

(ii) Suppose $k < m+\alpha \leqslant 2k+1/4$. We separate into two cases, $2k \leqslant m$, $2k > m$. If $2k \leqslant m$,

$$\begin{aligned} |E\hat{f}_1^{(2k)}(x) - f_n^{(2k)}(x)| &= \left| \int h_\sigma^{(2k)}(x-t) f_n(t) dt - f_n^{(2k)}(x) \right| \\ &= \left| \int f_n^{(2k)}(x-t) h_\sigma(t) dt - f_n^{(2k)}(x) \right| \\ &= \left| \int (f_n^{(2k)}(x-\sigma t) - f_n^{(2k)}(x)) h(t) dt \right| \\ &= O(1) \end{aligned}$$

so that

$$E\hat{f}_1^{(2k)}(x) = O(1). \qquad \dots \quad (3.11)$$

Also,

$$\operatorname{var}\{\hat{f}_1^{(2k)}(x)\} \leqslant \frac{1}{n_1} \int \{h_\sigma^{(2k)}(x-t)\}^2 f_n(t) dt$$

$$\leqslant \frac{1}{n_1} \|f_n\|_\infty \, \sigma^{-4k-1} \|h^{(2k)}\|_2^2. \qquad \dots \quad (3.12)$$

Then,

$$E\ \mathrm{var}\,(\hat{\theta}_{k2}\,|\,\hat{F}_1) \leqslant \frac{1}{n-n_1}\int E[\{f_1^{(2k)}(x)\}^2]f_n(x)dx$$
$$= O\,(n^{-2}\,\sigma^{-4k-1}+n^{-1})$$
$$= O\,(n^{-8(m+\alpha-k)/(1+4m+4\alpha)}). \qquad \ldots \quad (3.13)$$

If $2k > m$ we compute,

$$|E\hat{f}_1^{(2k)}(x)| = |\int h_\sigma^{(2k)}(x-t)f_n(t)dt|$$
$$= |\int h_\sigma^{(2k-m)}(x-t)f_n^{(m)}(t)dt|$$
$$= \sigma^{-2k+m}\,|\int h^{(2k-m)}(t)f_n^{(m)}(x-\sigma t)dt|$$
$$= \sigma^{-2k+m}\,|\int h^{(2k-m)}(t)\,\{f_n^{(m)}(x-\sigma t)-f_n^{(m)}(x)\}dt|$$
$$\leqslant g(x)\,\sigma^{m+\alpha-2k}\int|h^{(2k-m)})(t)|\,dt \qquad \ldots \quad (3.14)$$

Again, by (3.12) and (3.14)

$$E\ \mathrm{var}\,(\hat{\theta}_{k2}\,|\,\hat{F}_1) = O\,(n^{-2}\,\sigma^{-(4k+1)}+n^{-1}\,\sigma^{m+\alpha-2k})$$
$$= O\,(n^{-8(m+\alpha-k)(1+4m+4\alpha)}) \qquad \ldots \quad (3.15)$$

The result follows by (3.13), (3.15) and (3.9). $\square$

Proof of Theorem 2 : (i) Let $\{F_\nu\}$ be a sequence of distributions with densities f_ν and square root of densities s_ν. Suppose $\|s_\nu-s_0\|_2^2 \to 0$ and $\int\{f_\nu^{(2k)}(x)-f_0^{(2k)}(x)\}^2f_0(x)dx \to 0$.

Write, with some abuse of notation, $\theta_k(s_\nu) = \theta_k(F_\nu)$. Then,

$$\theta_k(s_\nu) = \int\{f_0^{(k)}(x)\}^2dx+2\int f_0^{(k)}(x)\{f_\nu^{(k)}(x)-f_0^{(k)}(x)\}dx+\int\{f_\nu^{(k)}(x)-f_0^{(k)}(x)\}^2\,dx. \qquad \ldots \quad (3.16)$$

Now

$$\int f_0^{(k)}(x)\,\{f_\nu^{(k)}(x)-f_0^{(k)}(x)\}dx = (-1)^k\int f_0^{(2k)}(x)f_\nu(x)dx-\theta_k(s_0)$$
$$= \int\{(-1)^kf_0^{(2k)}(x)-\theta_k(s_0)\}f_\nu(x)dx, \qquad \ldots \quad (3.17)$$

and

$$\int\{f_\nu^{(k)}(x)-f_0^{(k)}(x)\}^2dx$$
$$= (-1)^k\int\{f_\nu(x)-f_0(x)\}\,\{f_\nu^{(2k)}(x)-f_0^{(2k)}(x)\}dx$$
$$= (-1)^k\int\{s_\nu(x)-s_0(x)\}^2\,\{f_\nu^{(2k)}(x)-f_0^{(2k)}(x)\}dx$$
$$+2(-1)^k\int s_0(x)\{s_\nu(x)-s_0(x)\}\,\{f_\nu^{(2k)}(x)-f_0^{(2k)}(x)\}dx$$
$$\leqslant \|f_0^{(2k)}+f_\nu^{(2k)}\|_\infty\|s_\nu-s_0\|_2^2+2\|s_\nu-s_0\|_2\,[\int\{f_\nu^{(2k)}(x)-f_0^{(2k)}(x)\}^2f_0(x)dx]^{1/2}$$
$$= o(\|s_\nu-s_0\|_2). \qquad \ldots \quad (3.18)$$

(3.16), (3.17) and (3.18) imply that

$$\theta_k(s_\nu) = \theta_k(s_0)+2\int\{(-1)^kf_0^{(2k)}(x)-\theta_k(F_0)\}\,f_\nu(x)dx+O(\|s_\nu-s_0\|_2).$$

This means that $\theta_k(s)$ is Fréchet differentiable along such paths with derivative $4\{(-1)^k f_0^{(2k)}-\theta_k(F_0)\}s_0$ and the result follows by standard theory.

(ii) Here, as in Ritov and Bickel (1987) we prove the assertion by presenting a sequence of Bayes problems. In the nth problem we observe $X_1,\ldots,X_n$ iid, $X_1 \sim F\epsilon \boldsymbol{F}_{m,\alpha,g}$. The loss function is $\boldsymbol{L}_n(\theta, d) = 1_{\{|\theta-d|> c_n^{-1}n^{-\gamma}\}}$. F is picked according to a measure Π_ν to be described next. Note that the sequence $\Pi_1, \Pi_2,\ldots$ is constructed such that the union of their supports $\boldsymbol{F}^*$ is compact with F_0 its only accumulation point. Let $F_0\epsilon\boldsymbol{F}_{m,\alpha,g}$ be arbitrary. Clearly, f_0 is bounded away from zero on some interval. For simplicity we take this interval to be [0, 1]. To simplify the notation we assume also that $\sup\limits_{x\epsilon[0,1]} g(x) \leqslant 1$.

We now describe Π_ν. Let h_i, $i = 0, 1,\ldots, \nu-1$ be a sequence of functions such that $\int\limits_0^1 h_i(x)dx = 0$, $h_i^{(j)}(0) = h_i^{(j)}(1) = 0$, $j = 0,\ldots, m+1$, $\int\{h_i^{(k)}(x)\}^2\, dx = 1$ and $\int\limits_{i/\nu}^{(i+1)/\nu} h_i^{(k)}(\nu x-i)f_0^{(k)}(x)dx = 0$. Let β equal $0, 1, \ldots, r-1$ with probability $1/r$ and let $\Delta_0,\ldots, \Delta_{\nu-1}$ be iid, independent of β and each equal to ± 1 with probability 1/2. Let F be the random measure with density

$$f(x) = f_0(x)+\beta\nu^{-(m+\alpha)}\,\Delta_i h_i(\nu x-i) \text{ on } [i/\nu, (i+1)/\nu).$$

The measure that governs the selection of F is Π_ν. Clearly, for any F in the support of Π_ν by our assumptions of h_i,

$$\theta_k(F) = \theta_k(F_0)+\beta^2\nu^{-2(m+\alpha)+2k}.$$

That is $\theta_k(F)$ equals $\theta_k(F_0)+j\nu^{-2(m+\alpha-k)}$ if $\beta = j$.

We show that if

$$n^2\nu^{-(4m+4\alpha+1)}\rightarrow 0 \qquad \ldots \quad (3.19)$$

then the variational distance between the probability measures of $X_1, \ldots, X_n$ under $\beta = i$ and $\beta = j$ tends to 0. Assume that this is the case and F is distributed according to Π_ν, Π_ν satisfies (3.19) and

$$\nu^{-2(m+\alpha-k)}c_n n^\gamma \rightarrow \infty \qquad \ldots \quad (3.20)$$

where

$$\gamma = 4(m+\alpha-k)/(1+4m+4\alpha).$$

This is possible if $k < m+\alpha$. If

$$A_{nj} = \{|T_n-\theta_k(F_j)| < [c_n n^\gamma]^{-1}\}$$

then by construction for n sufficiently large the A_{nj} are disjoint. The Bayes risk for estimating $\theta_k(F)$ using our loss function is

$$R_n = \frac{1}{r}\sum_{j=1}^{r} P_j^{(n)}(A_{nj}^C)$$

$$= 1-\frac{1}{r}\sum_{j=1}^{r} P_j^{(n)}(A_{nj}).$$

But, by the equivalence of $P[\,.\,|\beta = i]$ and $P[\,.\,|\beta = j]$ we have observed

$$P_j^{(n)}(A_{nj})-P_0^{(n)}(A_{nj})\to 0 \text{ for each } j.$$

So,

$$\underline{\lim}_n R_n \geqslant 1-\frac{1}{r}\,\overline{\lim}\sum_{j=1}^{r} P_0^{(n)}(A_{nj})$$

$$= 1-\frac{1}{r}\,\overline{\lim}\, P_0^{(n)}\left(\bigcup_{j=1}^{r} A_{nj}\right) \geqslant 1-\frac{1}{r}.$$

Finally

$$\inf_{T_n}\sup_{F\epsilon F^*} P_F[c_n n^\gamma |T_n-\theta_k(F)| \geqslant 1] \geqslant R_n.$$

Hence, since r is arbitrary,

$$\underline{\lim}\inf_{T_n}\sup_{F\epsilon F^*} P_F[c_n n^\gamma |T_n-\theta_k(F)| \geqslant 1] = 1$$

as advertised. This combines ideas of Hasminskii (1979) and Stone (1983).

We turn to the proof that (3.22) implies convergence of the variational distance. Let N_i, $i = 0, \ldots, \nu-1$ be the number of X's in $[i/\nu, (i+1)/\nu)$ and let $X_{i1},\ldots, X_{iN_i}$ be the set of observations in that interval. Note that the random vector $(N_0, \ldots, N_{\nu-1})$ is independent of β and $(\Delta_0,\ldots,\Delta_{\nu-1})$, and that the blocks $(X_{i1}, \ldots, X_{iN_i})$ and $(X_{j1}, \ldots, X_{jN_j})$, $i \neq j$ are independent given N_i and N_j. Without loss of generality consider $\beta = 0$ and $\beta = 1$.

The likelihood ratio of $\beta = 1$ to $\beta = 0$ is $L = \prod_{i=0}^{\nu-1} L_i$ where

$$L_i = 1/2\prod_{j=1}^{N_i}\{1+\nu^{-(m+\alpha)}h_i(U_{ij})/f_0(U_{ij})\}+1/2\prod_{j=1}^{N_i}\{1-\nu^{-(m+\alpha)}h_i(U_{ij})/f_0(U_{ij})\}$$

$$= 1+\sum_{l=1}^{[N_i/2]}\nu^{-2(m+\alpha)l}\sum_{\substack{j_1,\ldots,j_{2l}\\ \text{all different}}}\cdots\sum \frac{h_i(U_{ij_1})}{f_0(U_{ij_1})}\cdots\frac{h_i(U_{ij_{2l}})}{f_0(U_{ij_{2l}})}$$

where $U_{ij} = \nu X_{ij}-i$ and $[x]$ is the greatest integer not larger than x.

Note that, $f_i(x) := \left[\nu \left\{ F_0 \left(\frac{i+1}{\nu}\right) - F_0 \left(\frac{i}{\nu}\right) \right\} \right]^{-1} f_0\left(\frac{i+x}{\nu}\right)$ is the density of U_{ij} under f_0. We show that $L \xrightarrow{P} 1$ under F_0, which implies that the variational distance between the two conditional distribution tends to 0.

Since $\int_0^1 h_i(x)dx = 0$,

$$E(L_i - 1 \mid N_i) = 0. \qquad \ldots (3.21)$$

Since $\|f_0\| < \infty$ by the lemma and the infimum of f_0 on $[0, 1]$ is > 0 by construction we obtain

$$\int_0^1 \frac{h_i^2(u)}{f_0^2(u)} f_i(u)du = \int_0^1 \frac{h_i^2(u)}{f_0(u)} \cdot \left[\nu\left\{F_0\left(\frac{i+1}{\nu}\right) - F_0\left(\frac{i}{\nu}\right)\right\}\right]^{-1} \leqslant \frac{1}{[\inf_{x\in[0,1]} f_0(x)]^2} < \infty.$$

Let $A = \sup_i \int_0^1 f_0^{-2}(u) f_i(u)\, h_i^2(u)\, du$. Then

$$\operatorname{var}(L_i - 1 \mid N_i) \leqslant \sum_{l=1}^{[N_i/2]} \nu^{-4(m+\alpha)l} \binom{N_i}{2l} A^{2l},$$

and

$$\operatorname{var}\left\{\sum_{i=0}^{\nu-1} (L_i - 1)\right\} = E\left\{\sum_{i=0}^{\nu-1} (L_i-1)^2\right\} \leqslant E \sum_{i=0}^{\nu-1} \sum_{l=1}^{[N_i/2]} \nu^{-4(m+\alpha)l} \binom{N_i}{2l} A^{2l}. \qquad \ldots (3.22)$$

Let $p_i = F_0((i+1)/\nu) - F_0(i/\nu)$. Straightforward calculations give

$$E \sum_{l=1}^{[N_i/2]} \nu^{-4(m+\alpha)l} \binom{N_i}{2l} A^{2l}$$

$$= \sum_{j=2}^{n} \binom{n}{j} p_i^j (1-p_i)^{n-j} \sum_{l=1}^{[j/2]} (A\nu^{-2(m+\alpha)})^{2l} \binom{j}{2l}$$

$$= \sum_{l=1}^{[n/2]} (A\nu^{-2(m+\alpha)})^{2l} \frac{n!}{(2l)!} \sum_{j=2l}^{n} \frac{1}{(j-2l)!\,(n-j)!} p_i^j (1-p_i)^{n-j}$$

$$= \sum_{l=1}^{[n/2]} (A\nu^{-2(m+\alpha)})^{2l} \binom{n}{2l} \sum_{j=0}^{n-2l} \frac{(n-2l)!}{j!\,(n-2l-j)!} p_i^{j+2l} (1-p_i)^{n-2l-j}$$

$$= \sum_{l=1}^{[n/2]} (A\, p_i\, \nu^{-2(m+\alpha)})^{2l} \binom{n}{2l}$$

$$\leqslant \sum_{l=1}^{[n/2]} \frac{1}{(2l)!} (nA\, p_i\, \nu^{-2(m+\alpha)})^{2l} \leqslant \exp\{nAp_i\, \nu^{-2(m+\alpha)}\}^2 - 1 \qquad \ldots (3.23)$$

$$= (1+o(1))A^2 n^2 p_i^2\, \nu^{-4(m+\alpha)} = O(n\, \nu^{-(2m+2\alpha+1)})^2$$

since $\nu\, p_i < \|f_0\|_\infty$.

We obtain from (3.22) and (3.23) that

$$\text{var}\left\{\sum_{i=0}^{\nu-1}(L_i-1)\right\} = O(n^{-2}\nu^{-(4m+4\alpha+1)})$$

Therefore, from (3.19) and (3.21) we obtain :

$$\sum_{i=0}^{\nu-1}(L_i-1) = o_P(1) \text{ and } \sum_{i=0}^{\nu-1}(L_i-1)^2 = o_P(1)$$

both under F_0. Hence

$$\log L = \sum_{i=0}^{\nu-1}(L_i-1)+O\left(\sum_{i=0}^{\nu-1}(L_i-1)^2\right) \xrightarrow{P} 0$$

under F_0 proving the assertion. □

Proof of Lemma 1: It is enough to prove that for any $\alpha_i > 0$ and $d_i < \infty$,

$$\sup_{\substack{x,y \\ 0<|x-y|\leqslant 1}} \{|f^{(i)}(x)-f^{(i)}(y)|/|x-y|^{\alpha_i}\} \leqslant d_i \qquad \dots \quad (3.24)$$

implies that

$$\|f^{(i)}\|_\infty \leqslant c_i \qquad \dots \quad (3.25)$$

where $c_i < \infty$ is a function of α_i and d_i only. Suppose (3.24) implies (3.25) then

$$|f^{(i-1)}(x)-f^{(i-1)}(y)| = |f^{(i)}(x^*)|\,|x-y| \leqslant c_i|x-y| \text{ for } 0 < |x-y| \leqslant 1$$

and the lemma follows by backward induction from m.

Suppose (3.1) holds. Let b_i be an arbitrary number lying in (0, 1] and assume that $f^{(i)}(x) \geqslant d_i(b_i/2)^{\alpha_i}$ for a point $x \in R$. Then

$$f^{(i)}(y) \geqslant a_i = f^{(i)}(x)-d_i(b_i/2)^{\alpha_i} \geqslant 0 \qquad \dots \quad (3.26)$$

for all $y \in [x-b_i/2, x+b_i/2] \equiv J_i$.

Then $f^{(i-1)}(u)$ is monotone on J_i and $|f^{(i-1)}(y)|$, $y \in J_i$, can be smaller than $a_{i-1} \equiv 1/4a_ib_i$ only on an interval of length smaller than $1/2b_i$. This leaves an interval J_{i-1} of length $b_{i-1} \geqslant 1/4b_i$ on which either $\inf_{y\in J_{i-1}} \{f^{(i-1)}(y)\} \geqslant a_{i-1}$ or $\sup_{y\in J_{i-1}} \{f^{(i-1)}(y)\} \leqslant -a_{i-1}$. Continue this line of argument inductively and obtain that (3.26) entails that $f(y) \geqslant a_0 \geqslant a_ib_i^i/2^{i(i+1)}$ on the interval J_0 whose length is $b_0 \geqslant 4^{-i}\, b_i$. But $f(\cdot)$ is a probability density function and hence

$$1 \geqslant a_0b_0 \geqslant 2^{-i(i+3)}\, a_i\, b_i^{i+1}.$$

Therefore,

$$f^{(i)}(x) = a_i + d_i(b_i/2)^{\alpha_i}$$
$$\leqslant 2^{i(i+3)} b_i^{-(i+1)} + d_i(b_i/2)^{\alpha_i}.$$

Hence $f^{(i)}$ is bounded and the lemma follows. □

Acknowledgment. P. Hall and S. Marron pointed out a gap in our original proof of Theorem 2 which we have corrected.

REFERENCES

BICKEL, P. J. (1982): On adaptive estimation. *Ann. Statist.*, **10**, 647-671.

DONOHO, D. L. and LIU, R. C. (1987): Geometrizing rates of convergence I–III (manuscript).

FARREL, R. H. (1972): On the best obtainable asymptotic rates of convergence of a density function at a point. *Ann. Math. Stat.*, **43**, 170-180.

HALL, P. and MARRON, J. S. (1987): Estimation of integrated squared density derivatives. To appear in *Statistical and Probability Letters.*

HASMINSKII, R. Z. (1979): Lower bound for the risks of nonparametric estimates of the mode. *Contributions to Statistics*, (J. Hájek Memorial Volume). *Academia, Prague*, 91-97.

HASMINSKII, R. Z. and IBRAGIMOV, I. A. (1978): On the non parametric estimation of functionals. *Symposium in Asymptotic Statistics, Prague*, 41-52.

IBRAGIMOV, I. A. and HASMINSKII, R. Z. (1981): *Statistical Estimation*, Springer Verlag, New York 237-240.

KHOSHEVNIK, YU, A. and LEVIT, B.YA. (1976): On a non-parametric analogue of the information matrix. *Theor. Probab. Appl.*, **21**, 738-753.

PFANZAGL, J. (1982): Contributions to a general asymptotic statistical theory. *Lecture Notes in Statistics*, **13**, Springer-Verlag, New York.

PRAKASA RAO, B. L. S. (1983): *Nonparametric Functional Estimation*, Academic Press, New York.

RITOV, Y. and BICKEL, P. J. (1987): Achieving information bounds in non and semi-parametric models. To appear in *Ann. Statist.*

SCHICK, A. S. (1986): On asymptotically efficient estimators in semiparametric models. *Ann. Statist.*, **14**, 1139-1151.

SCHWEDER, T. (1975): Window estimation of the asymptotic variance of rank estimators of location. *Scandinavian J. of Statist.*, **2**, 113-126.

STONE, C. J. (1983): Optimal uniform rate of convergence for nonparametric estimators of a density function and its derivatives. *Recent Advances in Statistics*, Paper in Honor of H. Chernoff. Academic Press, New York.

Paper received: September, 1988.

IMS Lecture Notes–Monograph Series
Complex Datasets and Inverse Problems: Tomography, Networks and Beyond
Vol. 54 (2007) 177–186

DOI: 10.1214/074921707000000148

Local polynomial regression on unknown manifolds

Peter J. Bickel[1] **and Bo Li**[2]

University of California, Berkeley and Tsinghua University

Abstract: We reveal the phenomenon that "naive" multivariate local polynomial regression can adapt to local smooth lower dimensional structure in the sense that it achieves the optimal convergence rate for nonparametric estimation of regression functions belonging to a Sobolev space when the predictor variables live on or close to a lower dimensional manifold.

1. Introduction

It is well known that worst case analysis of multivariate nonparametric regression procedures shows that performance deteriorates sharply as dimension increases. This is sometimes refered to as the curse of dimensionality. In particular, as initially demonstrated by [19, 20], if the regression function, $m(x)$, belongs to a Sobolev space with smoothness p, there is no nonparametric estimator that can achieve a faster convergence rate than $n^{-\frac{p}{2p+D}}$, where D is the dimensionality of the predictor vector X.

On the other hand, there has recently been a surge in research on identifying intrinsic low dimensional structure from a seemingly high dimensional source, see [1, 5, 15, 21] for instance. In these settings, it is assumed that the observed high-dimensional data are lying on a low dimensional smooth manifold. Examples of this situation are given in all of these papers — see also [14]. If we can estimate the manifold, we can expect that we should be able to construct procedures which perform as well as if we know the structure. Even if the low dimensional structure obtains only in a neighborhood of a point, estimation at that point should be governed by actual rather than ostensible dimension. In this paper, we shall study this situation in the context of nonparametric regression, assuming the predictor vector has a lower dimensional smooth structure. We shall demonstrate the somewhat surprising phenomenon, suggested by Bickel in his 2004 Rietz lecture, that the procedures used with the expectation that the ostensible dimension D is correct will, with appropriate adaptation not involving manifold estimation, achieve the optimal rate for manifold dimension d.

Bickel conjectured in his 2004 Rietz lecture that, in predicting Y from X on the basis of a training sample, one could automatically adapt to the possibility that the apparently high dimensional X that one observed, in fact, lived on a much smaller dimensional manifold and that the regression function was smooth on that manifold. The degree of adaptation here means that the worst case analyses for prediction are governed by smoothness of the function on the manifold and not on

[1]367 Evans Hall, Department of Statistics, University of California, Berkeley, CA, 94720-3860, USA, e-mail: bickel@stat.berkeley.edu

[2]S414 Weilun Hall, School of Economics and Management, Tsinghua University, Beijing, 100084, China, e-mail: libo@em.tsinghua.edu.cn

AMS 2000 subject classifications: primary 62G08, 62H12; secondary 62G20.

Keywords and phrases: local polynomial regression, manifolds.

the space in which X ostensibly dwells, and that purely data dependent procedures can be constructed which achieve the lower bounds in all cases.

In this paper, we make this statement precise with local polynomial regression. Local polynomial regression has been shown to be a useful nonparametric technique in various local modelling, see [8, 9]. We shall sketch in Section 2 that local linear regression achieves this phenomenon for local smoothness $p = 2$, and will also argue that our procedure attains the global IMSE if global smoothness is assumed. We shall also sketch how polynomial regression can achieve the appropriate higher rate if more smoothness is assumed.

A critical issue that needs to be faced is regularization since the correct choice of bandwidth will depend on the unknown local dimension $d(x)$. Equivalently, we need to adapt to $d(x)$. We apply local generalized cross validation, with the help of an estimate of $d(x)$ due to [14]. We discuss this issue in Section 3. Finally we give some simulations in Section 4.

A closely related technical report, [2] came to our attention while this paper was in preparation. Binev et al consider in a very general way, the construction of non-parametric estimation of regression where the predictor variables are distributed according to a fixed completely unknown distribution. In particular, although they did not consider this possibility, their method covers the case where the distribution of the predictor variables is concentrated on a manifold. However, their method is, for the moment, restricted to smoothness $p \leq 1$ and their criterion of performance is the integral of pointwise mean square error with respect to the underlying distribution of the variables. Their approach is based on a tree construction which implicitly estimates the underlying measure as well as the regression. Our discussion is considerably more restrictive by applying only to predictors taking values in a low dimensional manifold but more general in discussing estimation of the regression function at a point. Binev et al promise a further paper where functions of general Lipschitz order are considered.

Our point in this paper is mainly a philosophical one. We can unwittingly take advantage of low dimensional structure without knowing it. We do not give careful minimax arguments, but rather, partly out of laziness, employ the semi heuristic calculations present in much of the smoothing literature.

Here is our setup. Let $(X_i, Y_i), (i = 1, 2, \ldots, n)$ be i.i.d $\Re^{D+1}$ valued random vectors, where X is a D-dimensional predictor vector, Y is the corresponding univariate response variable. We aim to estimate the conditional mean $m_0(x) = E(Y|X = x)$ nonparametrically. Our crucial assumption is the existence of a local *chart*, i.e., each small patch of $\mathcal{X}$ (a neighborhood around x) is isomorphic to a ball in a d-dimensional Euclidean space, where $d = d(x) \leq D$ may vary with x. Since we fix our working point x, we will use d for the sake of simplicity. The same rule applies to other notations which may also depend on x.) More precisely, let $\mathcal{B}^d_{z,r}$ denote the ball in $\Re^d$, centered at z with radius r. A similar definition applies to $\mathcal{B}^D_{x,R}$. For small $R > 0$, we consider the neighborhood of x, $\mathcal{X}_x := \mathcal{B}^D_{x,R} \cap \mathcal{X}$ within $\mathcal{X}$. We suppose there is a continuously differentiable bijective map $\phi : \mathcal{B}^d_{0,r} \mapsto \mathcal{X}_x$. Under this assumption with $d < D$, the distribution of X degenerates in the sense that it does not have positive density around x with respect to Lebesgue measure on $\Re^D$. However, the induced measure $\mathbb{Q}$ on $\mathcal{B}^d_{0,r}$ defined below, can have a non-degenerate density with respect to Lebesgue measure on $\Re^d$. Let $\mathcal{S}$ be an open subset of $\mathcal{X}_x$, and $\phi^{-1}(\mathcal{S})$ be its preimage in $\mathcal{B}^{d(x)}_{0,r}$. Then $\mathbb{Q}(Z \in \phi^{-1}(\mathcal{S})) = \mathbb{P}(X \in \mathcal{S})$. We assume throughout that $\mathbb{Q}$ admits a continuous positive density function $f(\cdot)$. We proceed to our main result whose proof is given in the Appendix.

2. Local linear regression

[17] develop the general theory for multivariate local polynomial regression in the usual context, i.e., the predictor vector has a D dimensional compact support in $\Re^D$. We shall modify their proof to show the "naive" (brute-force) multivariate local linear regression achieves the "oracle" convergence rate for the function $m(\phi(z))$ on $\mathcal{B}^d_{0,r}$.

Local linear regression estimates the population regression function by $\hat{\alpha}$, where $(\hat{\alpha}, \hat{\beta})$ minimize

$$\sum_{i=1}^{n} \left(Y_i - \alpha - \beta^T(X_i - x)\right)^2 K_h(X_i - x).$$

Here $K_h(\cdot)$ is a $D-$variate kernel function. For the sake of simplicity, we choose the same bandwidth h for each coordinate. Let

$$X_x = \begin{bmatrix} 1 & (X_1 - x)^T \\ \vdots & \vdots \\ 1 & (X_n - x)^T \end{bmatrix}$$

and $W_x = diag\{K_h(X_1 - x), \ldots, K_h(X_n - x)\}$. Then the estimator of the regression function can be written as

$$\hat{m}(x, h) = e_1^T(X_x^T W_x X_x)^{-1} X_x^T W_x Y$$

where e_1 is the $(D+1) \times 1$ vector having 1 in the first entry and 0 elsewhere.

2.1. Decomposition of the conditional MSE

We enumerate the assumptions we need for establishing the main result. Let M be a canonical finite positive constant,

(i) The kernel function $K(\cdot)$ is continuous and radially symmetric, hence bounded.

(ii) There exists an $\epsilon(0 < \epsilon < 1)$ such that the following asymptotic irrelevance conditions hold.

$$E\Big[K^\gamma(\frac{X-x}{h})w(X)1\big(X \in (\mathcal{B}^D_{x,h^{1-\epsilon}} \cap \mathcal{X})^c\big)\Big] = o(h^{d+2})$$

for $\gamma = 1, 2$ and $|w(x)| \leq M(1 + |x|^2)$.

(iii) $v(x) = Var(Y|X = x) \leq M$.

(iv) The regression function $m(x)$ is twice differentiable, and $\|\frac{\partial^2 m}{\partial x_a x_b}\|_\infty \leq M$ for all $1 \leq a \leq b \leq D$ if $x = (x_1, \ldots, x_D)$.

(v) The density $f(\cdot)$ is continuously differentiable and strictly positive at 0 in $\mathcal{B}^d_{0,r}$.

Condition (ii) is satisfied if K has exponential tails since if $V = \frac{X-x}{h}$, the conditions can be written as

$$E\big[K^\gamma(V)w(x+hV)1(V \in (\mathcal{B}^D_{0,h^{1-\epsilon}})^c\big] = o(h^{d+2}).$$

Theorem 2.1. *Let x be an interior point in $\mathcal{X}$. Then under assumptions (i)-(v), there exist some $J_1(x)$ and $J_2(x)$ such that*

$$E\{\hat{m}(x,h) - m(x)|X_1,\dots,X_n\} = h^2 J_1(x)(1 + o_P(1)),$$

$$Var\{\hat{m}(x,h) - m(x)|X_1,\dots,X_n\} = n^{-1}h^{-d}J_2(x)(1 + o_P(1)).$$

Remark 1. The predictor vector doesn't need to lie on a perfect smooth manifold. The same conclusion still holds as long as the predictor vector is "close" to a smooth manifold. Here "close" means the noise will not affect the first order of our asymptotics. That is, we think of $X_1,\dots,X_n$ as being drawn from a probability distribution P on $\Re^D$ concentrated on the set

$$\mathcal{X} = \{y : |\phi(u) - y| \le \epsilon_n \text{ for some } u \in \mathcal{B}^d_{0,r}\}$$

and $\epsilon_n \to 0$ with n. It is easy to see from our arguments below that if $\epsilon_n = o(h)$, then our results still hold.

Remark 2. When the point of interest x is on the boundary of the support $\mathcal{X}$, we can show that the bias and variance have similar asymptotic expansions, following the Theorem 2.2 in [17]. But, given the extra complication of the embedding, the proof would be messier, and would not, we believe, add any insight. So we omit it.

2.2. Extensions

It's somewhat surprising but not hard to show that if we assume the regression function m to be p times differentiable with all partial derivatives of order p bounded ($p \ge 2$, an integer), we can construct estimates $\hat{m}$ such that,

$$E\{\hat{m}(x,h) - m(x)|X_1,\dots,X_n\} = h^p J_1(x)(1 + o_P(1)),$$

$$Var\{\hat{m}(x,h) - m(x)|X_1,\dots,X_n\} = n^{-1}h^{-d}J_2(x)(1 + o_P(1))$$

yielding the usual rate of $n^{-\frac{2p}{2p+d}}$ for the conditional MSE of $\hat{m}(x,h)$ if h is chosen optimal, $h = \lambda n^{-\frac{1}{2p+d}}$. This requires replacing local linear regression with local polynomial regression with a polynomial of order $p-1$. We do not need to estimate the manifold as we might expect since the rate at which the bias term goes to 0 is derived by first applying Taylor expansion with respect to the original predictor components, then obtaining the same rate in the lower dimensional space by a first order approximation of the manifold map. Essentially all we need is that, locally, the geodesic distance is roughly proportionate to the Euclidean distance.

3. Bandwidth selection

As usual this tells us, for $p = 2$, that we should use bandwidth $\lambda n^{-\frac{1}{4+d}}$ to achieve the best rate of $n^{-\frac{2}{4+d}}$. This requires knowledge of the local dimension as well as the usual difficult choice of λ. More generally, dropping the requirement that the bandwidth for all components be the same we need to estimate d and choose the constants corresponding to each component in a simple data determined way.

There is an enormous literature on bandwidth selection. There are three main approaches: plug-in ([7, 16, 18], etc); the bootstrap ([3, 11, 12], etc) and cross validation ([6, 10, 22], etc). The first has always seemed logically inconsistent to us since it requires higher order smoothness of m than is assumed and if this higher order smoothness holds we would not use linear regression but a higher order polynomial. See also the discussion of [23].

We propose to use a blockwise cross-validation procedure defined as follows. Let the data be $(X_i, Y_i), 1 \leq i \leq n$. We consider a block of data points $\{(X_j, Y_j) : j \in \mathcal{J}\}$, with $|\mathcal{J}| = n_1$. Assuming the covariates have been standardized, we choose the same bandwidth h for all the points and all coordinates within the block. A leave-one-out cross validation with respect to the block while using the whole data set is defined as following. For each $j \in \mathcal{J}$, let $\hat{m}_{-j,h}(X_j)$ be the estimated regression function (evaluated at X_j) via local linear regression with the whole data set except X_j. In contrast to the usual leave-one-out cross-validation procedure, our modified leave-one-out cross-validation criterion is defined as $mCV(h) = \frac{1}{n_1}\sum_{j\in\mathcal{J}}(Y_j - \hat{m}_{-j,h}(X_j))^2$. Using a result from [23], it can be shown that

$$mCV(h) = \frac{1}{n_1}\sum_{j\in\mathcal{J}} \frac{(Y_j - \hat{m}_h(X_j))^2}{(1 - S_h(j,j))^2}$$

where $S_h(j,j)$ is the diagonal element of the smoothing matrix S_h. We adopt the GCV idea proposed by [4] and replace the $S_h(j,j)$ by their average $atr_{\mathcal{J}}(S_h) = \frac{1}{n_1}\sum_{j\in\mathcal{J}} S_h(j,j)$. Thereby our modified generalized cross-validation criterion is,

$$mGCV(h) = \frac{1}{n_1}\sum_{j\in\mathcal{J}} \frac{(Y_j - \hat{m}_h(X_j))^2}{(1 - atr_{\mathcal{J}}(S_h))^2}.$$

The bandwidth h is chosen to minimize this criterion function.

We give some heuristics for the justifying the (blockwise homoscedastic) mGCV. In a manner analogous to [23], we can show

$$S_h(j,j) = e_1^T(X_x^T W_x X_x)^{-1} e_1 K_h(0)|_{x=X_j}.$$

In view of (A.2) in the Appendix, we see $S_h(j,j) = n^{-1}h^{-d}K(0)(A_1(X_j) + o_p(1))$. Thus as $n^{-1}h^{-d} \to 0$,

$$\begin{aligned} atr_{\mathcal{J}}(S_h) &= n^{-1}h^{-d}K(0)(n_1^{-1}\sum_{j\in\mathcal{J}} A_1(X_j) + o_p(1)) \\ &= O_p(n^{-1}h^{-d}) = o_p(1). \end{aligned}$$

Then, as is discussed in [22], using the approximation $(1-x)^{-2} \approx 1 + 2x$ for small x, we can rewrite $mGCV(h)$ as

$$mGCV(h) = \frac{1}{n_1}\sum_{j\in\mathcal{J}}(Y_j - \hat{m}_h(X_j))^2 + \frac{2}{n_1} tr_{\mathcal{J}}(S_h)\frac{1}{n_1}\sum_{j\in\mathcal{J}}(Y_j - \hat{m}_h(X_j))^2.$$

Now regarding $\frac{1}{n_1}\sum_{j\in\mathcal{J}}(Y_j - \hat{m}_h(X_j))^2$ in the second term as an estimator of the constant variance for the focused block, the mGCV is approximately the same as the C_p criterion, which is an estimator of the prediction error up to a constant.

In practice, we first use [14]'s approach to estimate the local dimension d, which yields a consistent estimate $\hat{d}$ of d. Based on the estimated intrinsic dimensionality $\hat{d}$, a set of candidate bandwidths $\mathcal{CB} = \{\lambda_1 n^{-\frac{1}{\hat{d}+4}}, \ldots, \lambda_B n^{-\frac{1}{\hat{d}+4}}\}$ $(\lambda_1 < \cdots < \lambda_B)$ are chosen . We pick the one minimizing the $mGCV(h)$ function.

4. Numerical experiments

The data generating process is as following. The predictor vector $X = (X_{(1)}, X_{(2)}, X_{(3)})$, where $X_{(1)}$ will be sampled from a standard normal distribution, $X_{(2)} = X_{(1)}^3 + sin(X_{(1)}) - 1$, and $X_{(3)} = \log(X_{(1)}^2 + 1) - X_{(1)}$. The regression function $m(x) = m(x_{(1)}, x_{(2)}, x_{(3)}) = cos(x^{(1)}) + x_{(2)} - x_{(3)}^2$. The response variable Y is generated via the mechanism $Y = m(X) + \varepsilon$, where ε has a standard normal distribution. By definition, the 3-dimensional regression function $m(x)$ is essentially a 1-dimensional function of $x_{(1)}$. $n = 200$ samples are drawn. The predictors are standardized before estimation. We estimate the regression function $m(x)$ by both the "oracle" univariate local linear (ull) regression with a single predictor $X_{(1)}$ and our blind 3-variate local linear regression with all predictors $X_{(1)}, X_{(2)}, X_{(3)}$.

We focus on the middle block with 100 data points, with the number of neighbor parameter k, needed for Levina and Bickel's estimate, set to be 15. The intrinsic dimension estimator is $\hat{d} = 1.023$, which is close to the true dimension, $d = 1$. We use the Epanechnikov kernel in our simulation. Our proposed modified GCV procedure is applied to both the ull and mll procedures. The estimation results are displayed in Figure 1. The $x - axis$ is the standardized $X_{(1)}$. From the right panel, we see the blind mll indeed performs almost as well as the "oracle" ull.

Next, we allow the predictor vector to only lie close to a manifold. Specifically, we sample $X_{(1)} = X'_{(1)} + \epsilon'_1, X_{(2)} = X'^3_{(1)} + sin(X'_{(1)}) - 1 + \epsilon'_2, X_{(3)} = \log(X'^2_{(1)} + 1) - X'_{(1)} + \epsilon'_3$, where $X'_{(1)}$ is sampled from a standard normal distribution, and ϵ'_1, ϵ'_2 and ϵ'_3 are sampled from $\mathcal{N}(0, \sigma'^2)$. The noise scale is hence governed by σ'. In our experiment, σ' is set to be $0.02, 0.04, \ldots, 0.18, 0.20$ respectively. The predictor vector samples are visualized in the left panel of Figure 2 with $\sigma' = 0.20$. In the maximum noise scale case, the pattern of the predictor vector is somewhat vague. Again, a blind "mll" estimation is done with respect to new data generated in the aforementioned way. We plot the MSEs associated with different noise scales in the right panel of Figure 2. The moderate noise scales we've considered indeed don't have a significant influence on the performance of the "mll" estimator in terms of MSE.

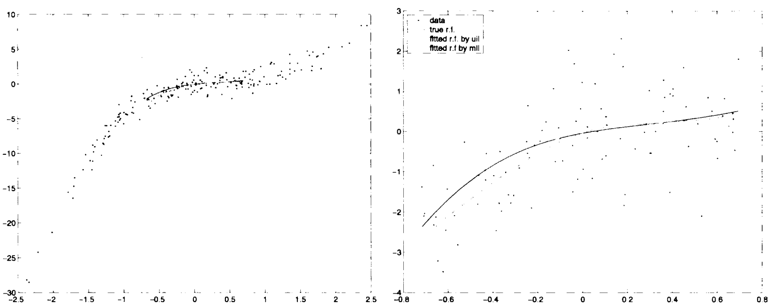


FIG 1. *The case with perfect embedding. The left panel shows the complete data and fitting of the middle block by both univariate local linear (ull) regression and multivariate local linear (mll) regression with bandwidths chosen via our modified GCV. The focused block is amplified in the right panel.*

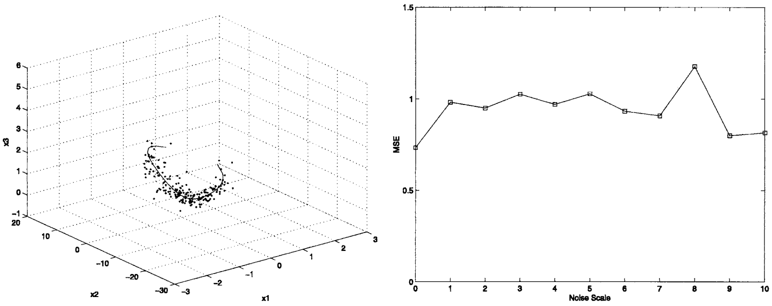


FIG 2. *The case with "imperfect" embedding. The left panel shows the predictor vector in a 3-D fashion with the noise scale* $\sigma' = 0.2$. *The right panel gives the MSEs with respect to increasing noise scales.*

Appendix

Proof of Theorem 2.1. Using the notation of [17], $\mathcal{H}_m(x)$ is the $D \times D$ Hessian matrix of $m(x)$ at x, and

$$Q_m(x) = [(X_1 - x)^T \mathcal{H}_m(x)(X_1 - x), \cdots, (X_n - x)^T \mathcal{H}_m(x)(X_n - x)]^T.$$

Ruppert and Wand have obtained the bias term.

$$\begin{aligned} &E(\hat{m}(x,h) - m(x)|X_1, \cdots, X_n) \\ &= \frac{1}{2} e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x \{Q_m(x) + R_m(x)\} \end{aligned} \tag{A.1}$$

where if $|\cdot|$ denotes Euclidean norm, $|R_m(x)|$ is of lower order than $|Q_m(x)|$. Also we have

$$\begin{aligned} &n^{-1} X_x^T W_x X_x \\ &= \begin{bmatrix} n^{-1}\sum_{i=1}^n K_h(X_i - x) & n^{-1}\sum_{i=1}^n K_h(X_i - x)(X_i - x)^T \\ n^{-1}\sum_{i=1}^n K_h(X_i - x)(X_i - x) & n^{-1}\sum_{i=1}^n K_h(X_i - x)(X_i - x)(X_i - x)^T \end{bmatrix}. \end{aligned}$$

The difference in our context lies in the following asymptotics.

$$\begin{aligned} EK_h(X_i - x) &= E\big[K_h(X_i - x) 1\big(X_i \in \mathcal{B}^D_{x,h^{1-\epsilon}} \cap \mathcal{X}\big)\big] \\ &\quad + E\big[K_h(X_i - x) 1\big(X_i \in \big(\mathcal{B}^D_{x,h^{1-\epsilon}} \cap \mathcal{X}\big)^c\big)\big] \\ &\overset{(ii)}{=} h^{-D}\Big(\int_{N^d_{0,h^{1-\epsilon}}} K\Big(\frac{\phi(z') - \phi(0)}{h}\Big) f(z') \mathrm{d}z' + o_P(h^d)\Big) \\ &= h^{d-D}\Big(f(0) \int_{\Re^d} K(\nabla\phi(0)u) \mathrm{d}u + o_P(1)\Big) \\ &= h^{d-D}\big(A_1(x) + o_P(1)\big). \end{aligned}$$

Thus, by the LLN, we have

$$n^{-1} \sum_{i=1}^n K_h(X_i - x) = h^{d-D}\big(A_1(x) + o_P(1)\big).$$

Similarly, there exist some $A_2(x)$ and $A_3(x)$ such that

$$n^{-1}\sum_{i=1}^{n} K_h(X_i - x)(X_i - x) = h^{2+d-D}\big(A_2(x) + o_P(1)\big)$$

and

$$n^{-1}\sum_{i=1}^{n} K_h(X_i - x)(X_i - x)(X_i - x)^T = h^{2+d-D}\big(A_3(x) + o_P(1)\big)$$

where we used assumption (i) to remove the term of order h^{1+d-D} in deriving the asymptotic behavior of $n^{-1}\sum_{i=1}^{n} K_h(X_i - x)(X_i - x)$. Invoking Woodbury's formula, as in the proof of Lemma 5.1 in [13], leads us to

$$\text{(A.2)} \qquad \big(n^{-1}X_x^T W_x X_x\big)^{-1} = h^{D-d}\begin{bmatrix} A_1(x)^{-1} + o_P(1) & O_P(1) \\ O_P(1) & h^{-2}O_p(1) \end{bmatrix}$$

On the other hand,

$$\begin{aligned} & n^{-1}X_x W_x Q_m(x) \\ &= \begin{bmatrix} n^{-1}\sum_{i=1}^{n} K_h(X_i - x)(X_i - x)^T \mathcal{H}_m(x)(X_i - x) \\ n^{-1}\sum_{i=1}^{n}\{K_h(X_i - x)(X_i - x)^T \mathcal{H}_m(x)(X_i - x)\}(X_i - x) \end{bmatrix}. \end{aligned}$$

In a similar fashion, we can deduce that for some $B_1(x), B_2(x)$,

$$n^{-1}\sum_{i=1}^{n} K_h(X_i - x)(X_i - x)^T \mathcal{H}_m(x)(X_i - x) = h^{2+d-D}\big(B_1(x) + o_P(1)\big)$$

and

$$n^{-1}\sum_{i=1}^{n}\{K_h(X_i - x)(X_i - x)^T \mathcal{H}_m(x)(X_i - x)\}(X_i - x) = h^{3+d-D}\big(B_2(x) + o_P(1)\big).$$

We have

$$\text{(A.3)} \qquad n^{-1}X_x W_x Q_m(x) = h^{d-D}\begin{bmatrix} h^2\big(B_1(x) + o_P(1)\big) \\ h^3\big(B_2(x) + o_P(1)\big) \end{bmatrix}.$$

It follows from (A.1),(A.2) and (A.3) that the bias admits the following approximation.

$$\text{(A.4)} \qquad E(\hat{m}(x,h) - m(x)|X_1, \dots, X_n) = h^2 A_1(x)^{-1}B_1(x) + o_P(h^2).$$

Next, we move to the variance term.

$$\text{(A.5)} \qquad \begin{aligned} & Var\{\hat{m}(x,h)|X_1, \dots, X_n\} \\ &= e_1^T(X_x^T W_x X_x)^{-1}X_x^T W_x V W_x X_x (X_x^T W_x X_x)^{-1} e_1. \end{aligned}$$

The upper-left entry of $n^{-1}X_x^T W_x V W_x X_x$ is

$$n^{-1}\sum_{i=1}^{n} K_h(X_i - x)^2 v(X_i) = h^{d-2D}C_1(x)(1 + o_P(1)).$$

The upper-right block is

$$n^{-1}\sum_{i=1}^{n} K_h(X_i - x)^2(X_i - x)^T v(X_i) = h^{1+d-2D}C_2(x)(1 + o_P(1))$$

and the lower-right block is

$$n^{-1}\sum_{i=1}^{n} K_h(X_i - x)^2 (X_i - x)(X_i - x)^T v(X_i) = h^{2+d-2D} C_3(x)(1 + o_P(1)).$$

In light of (A.2), we arrive at

$$Var\{\hat{m}(x,h)|X_1,\ldots,X_n\} = n^{-1}h^{-d}A_1(x)^{-2}C_1(x)(1 + o_P(1)). \tag{A.6}$$

The proof is complete. □

Acknowledgment. We thank Ya'acov Ritov for insightful comments.

References

[1] BELKIN, M. AND NIYOGI, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* **15** 1373–1396.

[2] BINEV, P., COHEN, A., DAHMEN, W., DEVORE, R. AND TEMLYAKOV, V. (2004). Universal algorithms for learning theory part i: piecewise constant functions. IMI technical reports, SCU.

[3] CAO-ABAD, R. (1991). Rate of convergence for the wild bootstrap in nonparametric regression. *Ann. Statist.* **19** 2226–2231. MR1135172

[4] CRAVEN, P. AND WAHBA, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31** 377–403. MR516581

[5] DONOHO, D. AND GRIMES, C. (2003). Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. USA* **100** 5591–5596 (electronic).

[6] DUDOIT, S. AND VAN DER LAAN, M. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat. Methodol.* **2** 131–154. MR2161394

[7] FAN, J. AND GIJBELS, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B* **57** 371–394. MR1323345

[8] FAN, J. AND GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London. MR1383587

[9] FAN, J. AND GIJBELS, I. (2000). Local polynomial fitting. In *Smoothing and Regression. Approaches, Computation and Application* (M.G. Schimek, ed.) 228–275. Wiley, New York.

[10] GYÖRFI, L., KOHLER, M., KRZYŻAK, A. AND WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York. MR1920390

[11] HALL, P., LAHIRI, S. AND TRUONG, Y. (1995). On bandwidth choice for density estimation with dependent data. *Ann. Statist.* **23** 2241–2263. MR1389873

[12] HÄRDLE, W. AND MAMMEN, E. (1991). Bootstrap methods in nonparametric regression. In *Nonparametric Functional Estmation and Related Topics (Spetses, 1990)* **335** 111–123. Kluwer Acad. Publ., Dordrecht. MR1154323

[13] LAFFERTY, J. AND WASSERMAN, L. (2005). Rodeo: Sparse nonparametric regression in high dimensions. Technical report, CMU.

[14] LEVINA, E. AND BICKEL, P. J. (2005). Maximum likelihood estimation of intrinsic dimension. *Advances in NIPS* **17**. MIT Press.

[15] Roweis, S. and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323–2326.

[16] Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Amer. Statist. Assoc.* **22** 1049–1062. MR1482136

[17] Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22** 1346–1370. MR1311979

[18] Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90** 1257–1270. MR1379468

[19] Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360. MR594650

[20] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR673642

[21] Tenenbaum, J. B., De Silva, V. and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.

[22] Wang, Y. (2004). Model selection. In *Handbook of Computational Statistics* 437–466. Springer, New York. MR2090150

[23] Zhang, C. (2003). Calibrating the degrees of freedom for automatic data smoothing and effective curve checking. *J. Amer. Statist. Assoc.* **98** 609–628. MR2011675

Chapter 5
Adaptive Estimation

Jon A. Wellner

5.1 Introduction to Four Papers on Semiparametric and Nonparametric Estimation

5.1.1 Introduction: Setting the Stage

I discuss four papers of Peter Bickel and coauthors: Bickel (1982), Bickel and Klaassen (1986), Bickel and Ritov (1987), and Ritov and Bickel (1990).

The four papers by Peter Bickel (and co-authors Chris Klaassen and Ya'acov Ritov) to be discussed here all deal with various aspects of estimation in semiparametric and nonparametric models. All four papers were published in the period 1982–1990, a time when semiparametric theory was in rapid development. Thus it might be useful to briefly review some of the key developments in statistical theory prior to 1982, the year in which Peter Bickel's Wald lectures (given in 1980) appeared, in order to give some relevant background information. Because I was personally involved in some of these developments in the early 1980s, my account will necessarily be rather subjective and incomplete. I apologize in advance for oversights and a possibly incomplete version of the history.

A key spur for the development of theory for semiparametric models was the clear recognition by Neyman and Scott (1948) that maximum likelihood estimators are often inconsistent in the presence of an unbounded (with sample size) number of nuisance parameters. The simplest of these examples is as follows: suppose that

$$(X_i, Y_i) \sim N_2((\mu_i, \mu_i), \sigma^2), \quad i = 1, \ldots, n \tag{5.1}$$

J.A. Wellner (✉)
Department of Statistics, University of Washington, Seattle, WA, USA
e-mail: jaw@stat.washington.edu

J. Fan et al. (eds.), *Selected Works of Peter J. Bickel*, Selected Works in Probability and Statistics 13, DOI 10.1007/978-1-4614-5544-8_5,

are independent where $\mu_i \in \mathbb{R}$ for $i = 1, \ldots, n$ and $\sigma^2 > 0$. Then the maximum likelihood estimator of σ^2 is

$$\hat{\sigma}_n^2 = (4n)^{-1} \sum_{i=1}^{n} (X_i - Y_i)^2 \to_p \frac{\sigma^2}{2}.$$

This is an example of what has come to be known as a "functional model". The corresponding "structural model" (or mixture or latent variable model) is: (X_i, Y_i) are i.i.d. with density $p_{\sigma,G}$ where

$$p_{\sigma,G}(x,y) = \int \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right) dG(\mu)$$

where ϕ is the standard normal density, $\sigma > 0$, and G is a (mixing) distribution on $\mathbb{R}$. Equivalently,

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} Z \\ Z \end{pmatrix} + \sigma \begin{pmatrix} \delta \\ \varepsilon \end{pmatrix}$$

where $Z \sim G$ is independent of $(\delta, \varepsilon) \sim N_2(0, I)$, and only (X, Y) is observed. Here the nuisance parameters $\{\mu_i, i = 1, \ldots, n\}$ of the functional model (5.1) have been replaced by the (nuisance) mixing distribution G. Kiefer and Wolfowitz (1956) studied general semiparametric models of this "structural" or mixture type, $\{p_{\theta,G} : \theta \in \Theta \subset \mathbb{R}^d,\ G \text{ a probability distribution}\}$, and established consistency of maximum likelihood estimators $(\hat{\theta}_n, \hat{G}_n)$ of (θ, G). (Further investigation of the properties of maximum likelihood estimators in structural models (or semiparametric mixture models) was pursued by Aad van der Vaart in the mid 1990s; I will return to this later.)

Nearly at the same time as the work by Kiefer and Wolfowitz (1956) and Stein (1956) studied efficient testing and estimation in problems with many nuisance parameters (or even nuisance functions) of a somewhat different type. In particular Stein considered the one-sample symmetric location model

$$\mathscr{P}_1 = \{p_{\theta,f}(x) = f(x-\theta) : \ \theta \in \mathbb{R}, \ f \text{ symmetric about } 0, \ I_f < \infty\}$$

and the two-sample (paired) shift model

$$\mathscr{P}_2 = \{p_{\mu,\nu,f}(x,y) = f(x-\mu)f(y-\nu) : \ \mu, \nu \in \mathbb{R}, \ I_f < \infty\};$$

here $I_f \equiv \int (f'/f)^2 f dx$. Stein (1956) studied testing and estimation in models $\mathscr{P}_1$ and $\mathscr{P}_2$, and established necessary conditions for "adaptive estimation": for example, conditions under which the information bounds for estimation of θ in the model $\mathscr{P}_1$ are the same as for the information bounds for estimation of θ in the sub-model in which f is known. Roughly speaking, these are both cases in which the efficient score and influence functions are orthogonal to the "nuisance tangent space" in $L_2^0(P)$; i.e. orthogonal to all possible score functions for regular parametric submodels for the infinite-dimensional part of the model. Models of this type, and in particular the symmetric location model $\mathscr{P}_1$, remained as a focus of research during the period 1956–1982.

Over the period 1956–1982, considerable effort was devoted to finding sufficient conditions for the construction of "adaptive estimators" and "adaptive tests" in the context of the model $\mathscr{P}_1$: Hájek (1962) gave conditions for the construction of adaptive tests in the model $\mathscr{P}_1$, while van Eeden (1970) gave a construction for the sub-model of $\mathscr{P}_1$ consisting of log-concave densities (for which the score function for location is monotone non-decreasing), Beran (1974) constructed efficient estimators based on ranks, while Stone (1975) gave a construction of efficient estimators based on an "estimated" one-step approach.

This, modulo a key paper by Efron (1977) on asymptotic efficiency of Cox's partial likelihood estimators, was roughly the state of affairs of semiparametric theory in 1980–1982. Of course this is an oversimplification: much progress had been underway from a more nonparametric perspective from several quarters: the group around Lucien Le Cam in Berkeley, including P. W. Millar and R. Beran, the Russian school including I. Ibragimov and R. Has'minskii in (now) St. Petersburg and Y. A. Koshevnik and B. Levit in Moscow, and J. Pfanzagl in Cologne. Over the decade from 1982 to 1993 these two directions would merge and be understood as a whole piece of cloth, but that was not yet the case in 1980–1982, the period when Peter Bickel gave his Wald Lectures (and prepared them for publication).

5.1.2 Paper 1

The first of these four papers, *On Adaptive Estimation*, represents the culmination and summary of the first period of research on the phenomena of adaptive estimation uncovered by Stein (1956): it gives a masterful exposition of the state of "adaptive estimation" in the early 1980s, and new constructions of efficient estimators in several models satisfying Stein's necessary conditions for "adaptive estimation" in the sense of Stein (1956). Bickel (1982) begins in Sect. 5.1.2 with an explanation of "adaptive estimation", with focus on the "i.i.d. case", and introduces four key examples to be treated: (1) the one-sample symmetric location model $\mathscr{P}_1$ introduced above; (2) linear regression with symmetric errors; (3) linear regression with a constant and arbitrary errors, a model closely related to the two-sample shift model $\mathscr{P}_2$ introduced above; and (4) location and variance-covariance parameters of elliptic distributions. The paper then moves to an explanation of Stein's necessary condition and presentation of a (new) set of sufficient conditions for adaptive estimation involving $L_2(P_{\theta_m,G})-$consistent estimation of the efficient influence function ("Condition H"). Bickel shows that the sufficient conditions are satisfied in the Examples (1)–(4), and hence that adaptive estimators exist in each of these problems. It was also conjectured that Condition H is necessary for adaptation. Necessary and sufficient conditions only slightly stronger than "Condition H" were established by Schick (1986) and Klaassen (1987); also see Bickel et al. (1993, 1998), Sect. 7.8.

According to the ISI Web of Science, as of 20 June 2011, this paper has received 228 citations, and thus is the most cited of the four papers reviewed here. It inspired the search for necessary and sufficient conditions for adaptive estimation (including the papers by Schick (1986) and Klaassen (1987) mentioned above). It also implicitly raised the issue of understanding efficient estimation in semiparametric models more generally. This was the focus of my joint work with Janet Begun, W. J. (Jack) Hall, and Wei-Min Huang at the University of Rochester during the period 1979–1983, resulting in Begun et al. (1983), which I will refer to in the rest of this discussion as BHHW.

5.1.3 Paper 2

Neyman and Scott (1948) had focused on inconsistency of maximum likelihood estimators in functional models, and Kiefer and Wolfowitz (1956) showed that inconsistency of likelihood-based procedures was not a difficulty for the corresponding structural (or mixture) models. Bickel and Klaassen (1986) initiated the exploration of efficiency issues in connection with functional models, with a primary focus on functional models connected with the symmetric location model $\mathscr{P}_1$. In particular, this paper examined the functional model with $X_i \sim N(\theta, \sigma_i^2)$ independent with $\sigma_i^2 \in \mathbb{R}^+$, $\theta \in \mathbb{R}$, for $1 \le i \le n$. The corresponding structural model is the normal scale mixture model with shift parameter θ, and hence is a subset of $\mathscr{P}_1$. In fact, it is a very rich subset with nuisance parameter tangent spaces (for "typical" points in the model) agreeing with that of the model $\mathscr{P}_1$. The main result of the paper is a theorem giving precise conditions under which a modified version of the estimator of Stone (1975) is asymptotically efficient, again in a precise sense defined in the paper.

This paper inspired further work on efficiency issues in functional models: see e.g. Pfanzagl (1993) and Strasser (1996). According to the ISI Web of Science (20 June 2011), it has been cited 15 times. These types of models remain popular (in September 2011, MathSciNet gives 414 hits for "functional model" and 480 hits for "structural model"), but many problems remain.

Between 1982 and publication of this paper in 1986, the paper Begun et al. (1983) appeared. In June 1983 Peter Bickel and myself had given a series of lectures at Johns Hopkins University on semiparametric theory as it stood at that time, and had started writing a book on the subject together with Klaassen and Ritov, Bickel et al. (1993, 1998), which was optimistically announced in the references for this paper as "BKRW (1987)".

5.1.4 Paper 3

This paper, Bickel and Ritov (1987), treats efficiency of estimation in the structural (or mixture model) version of the errors-in-variables model dating back at least to Neyman and Scott (1948) and Reiersol (1950), and perhaps earlier. As noted by the authors: "Estimates of β in the general Gaussian error model, with Σ_0 diagonal, have been proposed by a variety of authors including Neyman and Scott (1948) and Rubin (1956). In the arbitrary independent error model, Wolfowitz in a series of papers ending in 1957, Kiefer, Wolfowitz, and Spiegelman (1979) by a variety of methods gave estimates, which are consistent and in Spiegelman's case $n^{1/2}-$consistent and asymptotically. Little seems to be known about the efficiency of these procedures other than that in the restricted Gaussian model ...". This model is among the first semiparametric mixture models involving a nontrivial projection in the calculation of the efficient score function to receive a thorough analysis and constructions of asymptotically efficient estimators. The authors gave an explicit construction of estimators achieving the information bound in a very detailed analysis requiring 17 pages of careful argument.

The type of construction used by the authors involves kernel smoothing estimators of the nonparametric part of the model, and hence brings in choices of smoothing kernels and smoothing parameters (ε_n, c_n and v_n in the authors' notation, with $nc_n^2 v_n^6 \to \infty$). This same approach was used by van der Vaart (1988) to construct efficient estimators in a whole class of structural models of this same type; van der Vaart's construction involved the choice of seven different smoothing parameters. On the other hand, Pfanzagl (1990a) pages 47 and 48 (see also Pfanzagl 1990b) pointed out that the resulting estimators are rather artificial in some sense, and advocated in favor of maximum likelihood or other procedures requiring no (or at least fewer) smoothing parameter choices. This approach was pursued in van der Vaart (1996). Forty years after Kiefer and Wolfowitz established consistency of maximum likelihood procedures, Van der Vaart proved, efficiency of maximum likelihood in several particular structural models (under moment conditions which are sufficient but very likely not necessary), including the errors-in-variables model treated in the paper under review. The proofs in van der Vaart (1996) proceed via careful use of empirical process theory. Furthermore, Murphy and van der Vaart (1996) succeeded in extending the maximum likelihood estimators to confidence sets via profile likelihood considerations.

This paper has 35 citations in the ISI Web of Science as of 20 June 2011, but it inspired considerable further work on efficiency bounds and especially on alternative methods for construction of efficient estimators.

5.1.5 Paper 4

In the period 1988–1991 several key questions on the "boundary" between nonparametric and semiparametric estimation came under close examination by van der Vaart, Bickel and Ritov, and Donoho and Liu. The lower bound theory under development for publication in BKRW (1993) relied upon Hellinger differentiability of real-valued functionals. (The lower bound theory based on pathwise Hellinger differentiability was put in a very nice form by van der Vaart (1991).)

But the possibility of a gap between the conditions for differentiability and sufficient conditions to attain the bounds became a nagging question. In Ritov and Bickel (1990), Peter and Ya'acov analyzed the situation in complete detail for the real-valued functional $\nu(P) = \int p^2(x)dx$ defined for the collection $\mathscr{P}$ of distributions P on $[0,1]$ with a density p with respect to Lebesgue measure. This functional turns out to be Hellinger differentiable at all such densities p with an information lower bound given by

$$I_\nu^{-1} = 4Var(p(X)) = 4\int (p(x) - \nu(P))^2 p(x)dx.$$

However, Theorem 1 of Ritov and Bickel (1990) shows that there exist distributions $P \in \mathscr{P}$ such every sequence of estimators of $\nu(p)$ converges to $\nu(p)$ more slowly than $n^{-\alpha}$ for every $\alpha > 0$. It had earlier been shown by Ibragimov and Hasminskii (1979) that the $\sqrt{n}-$ convergence rate could be achieved for densities satisfying a Hölder condition of order at least $1/2$, and in a companion paper to the one under discussion Bickel and Ritov (1988), Peter and Ya'acov showed that this continued to hold for densities p satisfying a Hölder condition of at least $1/4$.

These results have been extended to obtain rates of convergence in the "non-regular" or nonparametric domain: see Birgé and Massart (1993, 1995) and Laurent and Massart (2000). More recently the techniques of analysis have been extended still further Tchetgen et al. (2008) and Robins et al. (2009). As of 20 June 2011, this paper has been cited 45 times (ISI Web of Science).

5.1.6 Summary and Further Problems

The four papers reviewed here represent only a small fraction of Peter Bickel's work on the theory of semiparametric models, but they illustrate his superb judgement in the choice of problems suited to push both the theory of semiparametric models in general terms and having relevance for applications. They also showcase his wonderful ability to see his way through the technicalities of problems to solutions of theoretical importance and which point the way forward to further understanding. Paper 1 was clearly important in development of general theory for the adaptive case beyond the location and shift models $\mathscr{P}_1$ and $\mathscr{P}_2$. Paper 2 initiated efficiency theory for estimation in functional models quite generally. Paper 3 played an important role in illustrating how semiparametric theory could be applied to the structural (or mixing) form of the classical errors in variables model, hence yielding one of the first substantial models to be discussed in detail in the "non-adaptive case" in which calculation of the efficient score and efficient influence function requires a non-trivial projection.

As noted by Kosorok (2009) semiparametric models continue to be of great interest because of their "... genuine scientific utility ... combined with the breadth and depth of the many theoretical questions that remain to be answered".

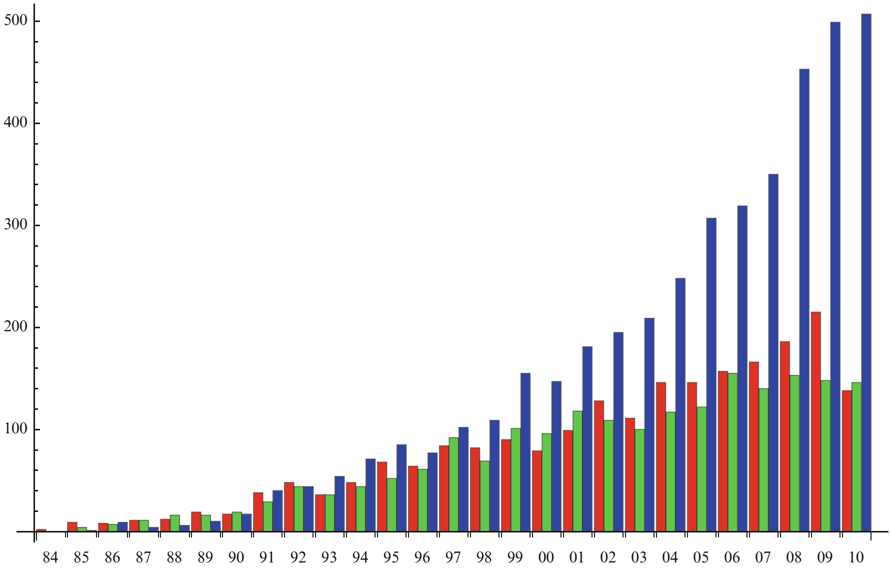


Fig. 5.1 Numbers of papers with "semiparametric" in title, keywords, or abstract, by year, 1984–2010. *Red* = MathSciNet; *Green* = Current Index of Statistics (CIS); *Blue* = ISI Web of Science

Figure 5.1 gives an update of Fig. 2.1 of Wellner et al. (2006). The trend is clearly increasing!

Acknowledgements Supported in part by NSF Grant DMS-0804587, and by NI-AID grant 2R01 AI291968-04.

References

Begun JM, Hall WJ, Huang W-M, Wellner JA (1983) Information and asymptotic efficiency in parametric–nonparametric models. Ann Stat 11(2):432–452

Beran R (1974) Asymptotically efficient adaptive rank estimates in location models. Ann Stat 2:63–74

Bickel PJ (1982) On adaptive estimation. Ann Stat 10(3):647–671

Bickel PJ, Klaassen CAJ (1986) Empirical Bayes estimation in functional and structural models, and uniformly adaptive estimation of location. Adv Appl Math 7(1):55–69

Bickel PJ, Ritov Y (1987) Efficient estimation in the errors in variables model. Ann Stat 15(2):513–540

Bickel PJ, Ritov Y (1988) Estimating integrated squared density derivatives: sharp best order of convergence estimates. Sankhyā Ser A 50(3):381–393

Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA (1993) Efficient and adaptive estimation for semiparametric models. Johns Hopkins series in the mathematical sciences. Johns Hopkins University Press, Baltimore

Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA (1998) Efficient and adaptive estimation for semiparametric models. Springer, New York. Reprint of the 1993 original
Birgé L, Massart P (1993) Rates of convergence for minimum contrast estimators. Probab Theory Relat Fields 97(1–2):113–150
Birgé L, Massart P (1995) Estimation of integral functionals of a density. Ann Stat 23(1):11–29
Efron B (1977) The efficiency of Cox's likelihood function for censored data. J Am Stat Assoc 72(359):557–565
Hájek J (1962) Asymptotically most powerful rank-order tests. Ann Math Stat 33:1124–1147
Ibragimov IA, Khasminskii RZ (1981) Statistical estimation: asymptotic theory. Springer Verlag, New York (Russian ed. 1979)
Kiefer J, Wolfowitz J (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann Math Stat 27:887–906
Klaassen CAJ (1987) Consistent estimation of the influence function of locally asymptotically linear estimators. Ann Stat 15(4):1548–1562
Kosorok MR (2009) What's so special about semiparametric methods? Sankhyā 71(2, Ser A):331–353
Laurent B, Massart P (2000) Adaptive estimation of a quadratic functional by model selection. Ann Stat 28(5):1302–1338
Murphy SA, van der Vaart AW (1996) Likelihood inference in the errors-in-variables model. J Multivar Anal 59(1):81–108
Neyman J, Scott EL (1948) Consistent estimates based on partially consistent observations. Econ 16:1–32
Pfanzagl J (1990a) Estimation in semiparametric models. Lecture notes in statistics, vol 63. Springer, New York. Some recent developments
Pfanzagl J (1990b) Large deviation probabilities for certain nonparametric maximum likelihood estimators. Ann Stat 18(4):1868–1877
Pfanzagl J (1993) Incidental versus random nuisance parameters. Ann Stat 21(4):1663–1691
Reiersol O (1950) Identifiability of a linear relation between variables which are subject to error. Econometrica 18:375–389
Ritov Y, Bickel PJ (1990) Achieving information bounds in non and semiparametric models. Ann Stat 18(2):925–938
Robins J, Tchetgen Tchetgen E, Li L, van der Vaart A (2009) Semiparametric minimax rates. Electron J Stat 3:1305–1321
Rubin H (1956) Uniform convergence of random functions with applications to statistics. Ann Math Statist 27:200–203
Schick A (1986) On asymptotically efficient estimation in semiparametric models. Ann Stat 14(3):1139–1151
Stein C (1956) Efficient nonparametric testing and estimation. In: Proceedings of the third Berkeley symposium on mathematical statistics and probability, 1954–1955, vol I. University of California Press, Berkeley/Los Angeles, pp 187–195
Stone CJ (1975) Adaptive maximum likelihood estimators of a location parameter. Ann Stat 3:267–284
Strasser H (1996) Asymptotic efficiency of estimates for models with incidental nuisance parameters. Ann Stat 24(2):879–901
Tchetgen E, Li L, Robins J, van der Vaart A (2008) Minimax estimation of the integral of a power of a density. Stat Probab Lett 78(18):3307–3311
van der Vaart AW (1988) Estimating a real parameter in a class of semiparametric models. Ann Stat 16(4):1450–1474
van der Vaart A (1991) On differentiable functionals. Ann Stat 19(1):178–204
van der Vaart A (1996) Efficient maximum likelihood estimation in semiparametric mixture models. Ann Stat 24(2):862–878
van Eeden C (1970) Efficiency-robust estimation of location. Ann Math Stat 41:172–181
Wellner JA, Klaassen CAJ, Ritov Y (2006) Semiparametric models: a review of progress since BKRW (1993). In: Frontiers in statistics. Imperial College Press, London, pp 25–44

The Annals of Statistics
1982, Vol 10, No. 3, 647–671

THE 1980 WALD MEMORIAL LECTURES

ON ADAPTIVE ESTIMATION

BY P. J. BICKEL[1]

University of California, Berkeley

We simplify a general heuristic necessary condition of Stein's for adaptive estimation of a Euclidean parameter in the presence of an infinite dimensional shape nuisance parameter and other Euclidean nuisance parameters. We derive sufficient conditions and apply them in the construction of adaptive estimates for the parameters of linear models and multivariate elliptic distributions. We conclude with a review of issues in adaptive estimation.

1. Introduction. In 1956, C. Stein published a paper in the Third Berkeley Symposium which deserves to be as well known as its celebrated companion piece on the inadmissibility of the normal mean. In this work Stein dealt with the problem of estimating and testing hypotheses about a Euclidean parameter θ or, more generally, a function $q(\theta)$ in the presence of an infinite dimensional "nuisance" shape parameter G. The question he asked (framed in estimation terms) was, "When can one estimate θ as well asymptotically not knowing G as knowing G?" He gave a simple necessary condition, which he checked in several important examples and, in one of these—testing that the center of symmetry has a specified value—he indicated a procedure that should work.

In recent years there has been considerable interest in an important situation where Stein's condition is satisfied, estimating the center of symmetry of an unknown symmetric distribution. Completely definitive results for this problem were obtained by Beran (1974) and Stone (1975). In this paper we return to Stein's original general formulation in the i.i.d. case. Motivated by his necessary condition for existence of adaptive estimates we obtain a simple sufficient condition for adaptation and apply it to a variety of important examples.

The paper is organized as follows. In Section 2 we define what we mean by adaptive estimation of θ; more precisely, we review some known results in the area and introduce the examples with which we will deal. In Section 3 we recall Stein's necessary condition for adaptation, and introduce a condition which we prove is sufficient. In Section 4 we check that our sufficient condition is satisfied in our examples. Section 5 contains a discussion of the connections between our work and recent research of Lindsay (1978, 1980), Hammerstrom (1978), Levitt (1974) and others, as well as a discussion of open questions. Finally, in Section 6, we gather technical parts of the proofs of our results.

2. What is adaptation? For simplicity we restrict ourselves throughout to the i.i.d. case. This is quite unnecessary for the heuristics of the paper. However, at least some of our proofs employ the assumed independence of the observations quite heavily.

Let $X_1, \cdots, X_n$ be i.i.d. k dimensional vectors with common distribution F. Let us recall the basic facts about the asymptotic theory of estimation when F ranges over a parametric model as put into their most elegant form by Le Cam.

Suppose that F is of the form F_θ where $\theta \in \Theta$, an open subset of R^p, and the F_θ have densities which we denote by $f(\cdot, \theta)$ with respect to a sigma-finite measure μ on R^k. Write

---

Received March 1981; revised November 1981.

[1] Research partially supported by ONR Grant No. N00014-80-C-0163 and the Adolph C. and Mary Sprague Miller Foundation for Basic Research in Science.

AMS 1970 *subject classification.* 62F20, 62G20.

Key words and phrases. Adaptation, efficient estimation, linear models, elliptic distributions.

E_θ, P_θ, $\mathscr{L}_\theta$ respectively for expectations, probabilities, and laws when θ holds. Let $\ell(x, \theta) = \log f(x, \theta)$, and define the following regularity conditions.

CONDITIONS R. *For all* $\theta \in \Theta$,

(i) $\ell(\cdot, \theta)$ *is differentiable in (the components of)* θ a.e. P_θ *and* $\dot\ell = (\partial\ell/\partial\theta_1, \cdots, \partial\ell/\partial\theta_p)$.
(ii) *The Fisher information matrix* $I(\theta)$ *exists*, $I(\theta) = E_\theta\{\dot\ell^T\dot\ell\,(X_1, \theta)\} < \infty$;
(iii) *Square root likelihood is differentiable in quadratic means, i.e. as* $t \to 0$,

$$E_\theta\left[\left\{\frac{f(X_1, \theta + t)}{f(X_1, \theta)}\right\}^{1/2} - 1 - \frac{t}{2}\dot\ell^T(X_1, \theta)\right]^2 = o(|t|^2),$$

and

$$P_{\theta+t}\{f(X_1, \theta) = 0\} = o(|t|^2),$$

where $|\cdot|$ *denotes the Euclidean norm (cf.* b_1 *and* b_2 *on page* 10 *of Le Cam*, 1969).
(iv) *There exist* $n^{1/2}$*-consistent estimates of* θ, *i.e.* $\{\tilde\theta_n(X_1, \cdots, X_n)\}$ *such that* $n^{1/2}(\tilde\theta_n - \theta) = O_{P_\theta}(1)$.

Under these conditions the following theorem holds (Le Cam, 1969; Fabian and Hannan, 1980). Call θ a *regular* point if $I(\theta)$ is nonsingular and if $I(\cdot)$ is continuous at θ.

THEOREM 2.1. *Under Conditions* R *there exist estimates* $\{\hat\theta_n\}$ *such that*

(a) *For all regular,* θ, $\mathscr{L}_{\theta_n}\{n^{1/2}(\hat\theta_n - \theta_n)\} \to \mathscr{N}(0, I^{-1}(\theta))$ *whenever* $n^{1/2}|\theta_n - \theta| \le M$ *for all* n, $M < \infty$.
(b) *The estimates* $\{\hat\theta_n\}$ *are asymptotically locally sufficient in the sense of Le Cam* (1969) *and locally asymptotically minimax in the sense of Hájek* (1972) *as modified by Fabian and Hannan* (1980).

Statement (a) says that $\{\hat\theta_n\}$ are efficient in the usual sense. Hájek (1972) also establishes, for $k = 1$, that any estimates satisfying (a) also are efficient in the sense of Rao. That is, if we define $\Delta_n(\cdot)$ by

(2.1) $$\hat\theta_n = \theta + n^{-1}\textstyle\sum_{i=1}^n \dot\ell(X_i, \theta)I^{-1}(\theta) + \Delta_n(\theta),$$

then

(2.2) $$n^{1/2}\Delta_n(\theta) \to_{P_\theta} 0,$$

for θ_n as in the theorem. In Theorem 6.1 (Section 6.4) we extend this result to general k.

REMARK 1. The construction of $\hat\theta_n$ used by Le Cam will prove useful to us later. Let $R_n^k = \{n^{-1/2}(i_1, \cdots, i_k), i_1, \cdots, i_k$ are arbitrary integers$\}$, and let

(2.3) $$\bar\theta_n = \text{the point in } R_n^k \text{ closest to } \tilde\theta_n.$$

If $\dot\ell^*(x, \theta)$ has the property that

$$n^{-1/2}\textstyle\sum_{i=1}^n \{\dot\ell^*(X_i, \theta_n) - \dot\ell(X_i, \theta)\} + n^{1/2}(\theta_n - \theta)I(\theta) = o_{P_\theta}(1)$$

whenever $n^{1/2}|\theta_n - \theta| \le M$, then Theorem 4 of Le Cam (1969) shows that

(2.4) $$\hat\theta_n = \bar\theta_n + n^{-1}\textstyle\sum_{j=1}^n \dot\ell^*(X_j, \bar\theta_n)I^-(\bar\theta_n)$$

is efficient in the sense of Theorem 2.1; where I^- is a generalized inverse of I. Of course, this construction is not unique and has unpleasant aspects such as the "discretization" of $\tilde\theta_n$ and its non-iterative character. However, the construction works in great generality, i.e., under the mild and natural Conditions R(i)–R(iv).

We shall actually want to take $\dot\ell^* = \dot\ell$. To do so we need an inconsequential strengthening of R(iii) which is valid in all our examples. We call UR(iii) the assumption that for all θ

$\in \Theta$, the differentiability condition of R(iii) holds uniformly in some neighbourhood of θ. We show in Theorem 6.2 (Section 6.4) that R(i), R(ii) and UR(iii) enable us to take $\dot{\ell}^* = \dot{\ell}$ in (2.4).

REMARK 2. Condition R(iv), although clearly necessary, appears hard to verify. In fact, Le Cam shows that if we assume identifiability of θ and nonsingularity of $I(\theta)$ for all $\theta \in \Theta$, R(i)-R(iii) imply R(iv). We have chosen to leave R(iv) in its present form for reasons which will be apparent later.

In a preprint which we saw after our lectures were prepared, Fabian and Hannan (1980) give a very careful treatment of estimation in locally asymptotically normal families. They present, among other results, the "right" version of Hájek's local asymptotic minimaxity, as well as a rigorous discussion of Stein's (1956) necessary conditions for adaptation. Their notion of adaptation agrees with ours (in their more general framework).

The models for which we will discuss adaptation may be described as follows: The common d.f. F of the X_i ranges over a set which can be parametrized by a Euclidean parameter θ of interest, and a shape nuisance parameter G, i.e.,

$$\mathscr{F} = \{F_{(\theta, G)} : \theta \in \Theta,\ G \in \mathscr{G}\} \tag{2.5}$$

where Θ is an open subset of R^p, $\mathscr{G}$ is a set of distributions on some space, and the map $(\theta, G) \to F_{(\theta, G)}$ is known.

For each $G \in \mathscr{G}$, define

$$\mathscr{F}_G = \{F_{(\theta, G)} : \theta \in \Theta\}. \tag{2.6}$$

The models $\mathscr{F}_G$ are parametric models. Suppose that $\mathscr{F}_G$ satisfies R(i), R(ii) and UR(iii) for each $G \in \mathscr{G}$. Define $f(\cdot, \theta, G)$, $\ell(\cdot, \theta, G)$, $I(\theta, G)$ respectively as density, log likelihood, and information in $\mathscr{F}_G$. Call (θ, G) *regular* if θ is regular in $\mathscr{F}_G$. Finally, in view of the Le Cam theorem, we can state the following definition.

DEFINITION. *A sequence of estimates $\{\hat{\theta}_n\}$ is adaptive if and only if, for every regular (θ, G),*

$$\mathscr{L}_{\theta_n}\{n^{1/2}(\hat{\theta}_n - \theta_n)\} \to \mathscr{N}(0, I^{-1}(\theta, G)) \tag{2.7}$$

whenever $n^{1/2} | \theta_n - \theta |$ stays bounded. Thus adaptive estimates, if they exist, are efficient for every $\mathscr{F}_G$ even though knowledge of the true G may not be used in the construction of the estimates.

Adaptive estimates of θ have been constructed in the first of our examples.

EXAMPLE 1. *Estimation of the center of symmetry.* Let $k = p = 1$. Take $\Theta = R$, $\mathscr{G} = \{$All distributions symmetric about 0$\}$, $F_{(\theta, G)}(x) = G(x - \theta)$.

The problem of adaptive estimation of θ in this model began to be studied by van Eeden (1970) and Takeuchi (1971), although the corresponding testing problem was earlier considered by Stein (1956) and solved by Hájek (1962). The definitive theorem was obtained by Beran (1974) and Stone (1975).

Let

$$I(G) = \int \{g'(x)\}^2 / g(x)\, dx \tag{2.8}$$

whenever g, the density of G, is absolutely continuous, and let $I(G) = \infty$ otherwise.

THEOREM 2.2. *There exist translation and scale equivariant estimates, $\{\hat{\theta}_n\}$ such that*

$$\mathscr{L}_{(0, G)}(n^{1/2}\hat{\theta}_n) \to \mathscr{N}(0, I^{-1}(G)) \tag{2.9}$$

for all $G \in \mathscr{G}$ *with* $I(G) < \infty$.

Hájek (1962) has shown that for this model (θ, G) is regular if $I(\theta, G) = I(G) < \infty$. The converse is also true. Thus $\{\hat{\theta}_n\}$ are adaptive according to our general definition. In fact, Stone (1975) shows that the estimates he constructs satisfy (2.9) with $I^{-1}(G) = 0$ whenever $I(G) = \infty$. □

We will construct adaptive estimates of θ in the following generalization of Example 1.

EXAMPLE 2. *Estimation of regression with symmetric errors.* We describe the model structurally in terms of a variable $X \sim F_{(\theta, G)}$. Here $k = p + 1$ and $\Theta = R^p$. Let

$$X = (C, Y) \tag{2.10}$$

where C is a p dimensional random vector and Y a scalar. Further,

$$Y = C\theta^T + \varepsilon \tag{2.11}$$

where $\varepsilon \sim G$, and ε and C are independent. We again take

$$\mathscr{G} = \{\text{All distributions } G \text{ on } R \text{ symmetric about } 0\}.$$

Finally, we suppose

$$E(C^T C) \text{ is nonsingular.} \tag{2.12}$$

This is just a stochastic version of the usual multiple regression model,

$$X_i = C_i\theta^T + \varepsilon_i, \qquad i = 1, \cdots, n,$$

where $C_1, \cdots, C_n$ are p dimensional vectors of constants such that $C^T = (C_1^T, \cdots, C_n^T)$ and $C^T C$ is nonsingular.

We deliberately do not specify that the distribution of C is known. The adaptive estimates we construct depend only on the data and work for any distribution of C satisfying (2.12). □

In many interesting situations a parameter θ for which efficient estimates exist in every model $\mathscr{F}_G$ cannot be consistently estimated in $\mathscr{F}$ because the parameter becomes unidentifiable. This is true in the next two examples. However, in both, natural functions $q(\theta)$ can be so estimated. In fact, adaptive estimation of these functions is possible. The definition of adaptive estimation of q is straightforward:

DEFINITION. *Suppose* $q : \Theta \to R^d$, $d \le p$, *has a total differential* $\dot{q}(\theta)$, *a* $d \times p$ *matrix. A sequence of estimates* $\{\hat{q}_n\}$ *of* q *is adaptive if and only if, for every regular* (θ, G),

$$\mathscr{L}_{\theta_n}\{n^{1/2}(q_n - q(\theta_n))\} \to \mathscr{N}(0, \dot{q}(\theta)I^{-1}(\theta, G)\dot{q}(\theta)^T) \tag{2.13}$$

whenever $n^{1/2} \mid \theta_n - \theta \mid$ *stays bounded.*

EXAMPLE 3. *Regression with a constant and arbitrary errors.* In Example 2, let $C = (C^\circ, 1)$, C° a $p - 1$ dimensional vector. Define X, Y, ε as before and suppose ε and C are independent. However, let $\mathscr{G} = \{\text{all distributions on } R\}$, and replace (2.12) by

$$E(C^\circ - EC^\circ)^T(C^\circ - EC^\circ) \quad \text{nonsingular.} \tag{2.14}$$

Evidently θ is not identifiable in $\mathscr{F}$ since a change in the constant θ_p could equally well be a change in G. However, $q(\theta) = (\theta_1, \cdots, \theta_{p-1})$ can be adaptively estimated, as we shall see.

A special case of this model, where $p = 2$ and

$$C^\circ = \begin{cases} 1 & \text{with probability } \lambda \\ 0 & \text{with probability } 1 - \lambda, \end{cases}$$

can be thought of as a two-sample model with random sample sizes, i.e., we observe N observations with distribution $G(x - \theta_1 - \theta_2)$ and $n - N$ observations with distribution $G(x - \theta_2)$, where N has a binomial (n, λ) distribution.

Adaptation in the two-sample model with fixed sample sizes (and unknown scale) was studied by Stein (1956), Weiss and Wolfowitz (1970), and Wolfowitz (1974). A definitive result was obtained by Beran (1974). Weiss and Wolfowitz (1971) considered the fixed sample size multiple regression model and obtained partial results. □

EXAMPLE 4. *Parameters of elliptic distributions.* The following multivariate generalization of the symmetric one-sample location and scale model has been considered by Huber (1977) and others. Let

$$X = \mu + \varepsilon V^{-1/2}$$

where μ is an unknown $1 \times k$ vector, V is a positive definite $k \times k$ symmetric matrix, and $V^{-1/2}$ is the unique positive definite symmetric square root of V^{-1}. We suppose $\varepsilon \sim G$, where

$$\mathscr{G} = \{G : G \text{ absolutely continuous, spherically symmetric on } R^k\}.$$

Take $\theta = (\mu, [V])$ where for any symmetric $k \times k$ matrix $M = \| m_{ij} \|$, we define $[M]$ to be the lexicographically written row vector of the lower $k(k+1)/2$ entries of M. Thus, $p = k(k+3)/2$ and

$$\Theta = \{(\mu, [V]) : V \text{ symmetric positive definite}\}$$

is an open subset of R^p.

Here θ is efficiently estimable at regular points of $\mathscr{F}_G$ but is not identifiable in $\mathscr{F}$. A common scale change in all coordinates is ascribable to either V or G, yet $(\mu, V/\text{tr } V)$ can be estimated consistently, in fact, adaptively, as we shall see.

3. Stein's considerations and a sufficient condition for adaptation. We begin by recalling Stein's necessary condition for adaptation. Define a parametric subfamily of $\mathscr{G}$ as a set $\{\mathscr{G}_\eta\}$, $\eta \in T$, where T is an open set in R^t and the map $\eta \to G_\eta$ is smooth. The parametric submodel of $\mathscr{F}$ corresponding to the parametric subfamily $\{G_\eta\}$ is naturally defined by $\{F_{(\theta, G_\eta)} : \theta \in \Theta, \eta \in T\}$. Here is Stein's necessary condition.

CONDITION S. *For every parametric submodel obeying* R(i)–R(iv) *with* $G_{\eta_0} = G_0$

$$\int \left\{ \frac{\partial}{\partial \theta_i} \ell(x, \theta, G_\eta) \frac{\partial}{\partial \eta_j} \ell(x, \theta, G_\eta) \right\}_{\theta = \theta_0, \eta = \eta_0} f(x, \theta_0, G_0) \mu(dx) = 0 \tag{3.1}$$

$$i = 1, \cdots, p, \quad j = 1, \cdots, t.$$

Stein (1956) shows that if an adaptive estimate of θ exists and (θ_0, G_0) is regular, then Condition S must hold. The argument is simple. Let

$$I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix},$$

where I_{11} is $p \times p$ and I_{22} is $t \times t$, be the $(p + t) \times (p + t)$-dimensional Fisher information matrix of the parametric submodel $F_{(\theta, G_\eta)}$ evaluated at (θ_0, η_0), and write

$$I^{-1} = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix}.$$

Now, by definition, if $\{\hat\theta_n\}$ is adaptive, then $I_{11}^{-1} = I^{-1}(\theta_0, G_0)$ is the asymptotic variance covariance matrix of $n^{1/2}(\hat\theta_n - \theta_n)$ whenever $n^{1/2} | \theta_n - \theta |$ stays bounded. But, by Hájek's (1972) theorem, I^{11} is the smallest variance covariance matrix achievable in this way. Thus $I_{11}^{-1} = I^{11}$ which is equivalent to $I_{12} = 0$, which is Condition S.

Condition S suffers from two defects: (i) it can be awkward to verify, (ii) it is unclear how to proceed from it to the construction of adaptive procedures. We now proceed to derive a simpler condition which is at least heuristically necessary and which in turn leads to a verifiable sufficient condition.

All the examples we have studied exhibit the following simple convexity structure:

CONDITION C. *$\mathscr{G}$ is convex and $G_0, G_1 \in \mathscr{G}$ implies that for $0 \le \alpha \le 1$*

$$F_{(\theta,\alpha G_0+(1-\alpha)G_1)} = \alpha F_{(\theta,G_0)} + (1-\alpha)F_{(\theta,G_1)}.$$

This structure suggests that we examine Condition S for the following $\{G_\eta\}$. Fix G_0 and G_1, take $T = (0, 1)$, and let

$$G_\eta = \eta G_0 + (1-\eta)G_1.$$

Then Condition S becomes for $\eta > 0$, $i = 1, \cdots, p$,

$$\int \frac{\partial}{\partial\theta_i} \ell(x, \theta, G_\eta)\{f(x, \theta, G_1) - f(x, \theta, G_0)\}\mu(dx) = 0.$$

Letting $\eta \to 0$ formally we get for "all" $G_0, G_1 \in \mathscr{G}$ that the following holds.

CONDITION S*.

$$\int \dot{\ell}(x, \theta, G_0)f(x, \theta, G_1)\mu(dx) = 0.$$

It may be shown formally that if Condition S* holds, so does Condition S (Bickel, 1979). Condition S* has a simple heuristic interpretation. If G_0 is a fixed shape in $\mathscr{G}$ let θ_n^* be the M-estimate corresponding to G_0, i.e., solving

$$\textstyle\sum_{i=1}^n \dot{\ell}(x_i, \theta_n^*, G_0) = 0.$$

We know that, under regularity conditions (Huber, 1967), if Condition S* holds, then $n^{1/2}(\theta_n^* - \theta)$ is asymptotically normal under $F_{(\theta,G)}$ with mean 0 and variance covariance matrix $A^{-1}B(A^T)^{-1}$, where

$$(3.2)\qquad \begin{aligned} A &= \left\| -\int \frac{\partial^2}{\partial\theta_i\partial\theta_j} \ell(x, \theta, G_0)f(x, \theta, G)\mu(dx)\right\|, \\ B &= \int \dot{\ell}^T(x, \theta, G_0)\dot{\ell}(x, \theta, G_0)f(x, \theta, G)\mu(dx). \end{aligned}$$

A heuristic summary of this is as follows. Firstly, M-estimates corresponding to a fixed shape G_0 should be $n^{-1/2}$ consistent for θ under every shape G_1. Secondly, suppose we can estimate the true G by data-dependent $\{G_n\}$ so that the score functions $\ell(\cdot, \cdot, G_n)$ converge to $\ell(\cdot,\cdot, G)$ and so that the matrices A_n, B_n obtained by replacing G_0 by G_n in (3.2) converge to $I(\theta, G)$. It then seems plausible that the sequence of M-estimates corresponding to G_n is adaptive.

Motivated by these considerations we now formulate two conditions, GR(iv) and H.

CONDITION GR(iv). *There exist estimates $\{\tilde{\theta}_n\}$ such that $n^{1/2}(\tilde{\theta}_n - \theta) = O_{P_{(\theta,G)}}(1)$ at all regular points (θ, G).*

Let

$$(3.3)\qquad \begin{aligned} \mathscr{H} = \{h: h \quad &\text{maps} \quad R^k \times \Theta \quad \text{to} \quad R^k \quad \text{and} \\ &\int h(x, \theta)F_{(\theta,G)}(dx) = 0 \quad \text{for all} \quad \theta \in \Theta, G \in \mathscr{G}\}. \end{aligned}$$

In view of Condition S*, $\mathscr{H}$ includes the space of possible score functions. For convenience we introduce

$$\tilde{\ell}(x, \theta, G) = \dot{\ell}(x, \theta, G)I^{-}(\theta, G), \tag{3.4}$$

where I^{-} is any generalized inverse. (In fact we only need $\tilde{\ell}$ for θ such that $I(\theta, G)$ is nonsingular.) Note that $\tilde{\ell}$ can be substituted for $\dot{\ell}$ in Condition S*. Here is our main condition:

CONDITION H. *Appropriate consistent estimation of score functions is possible. That is, there exists a sequence of maps* $\hat{\hat{\ell}}_m: (R^k)^m \to \mathscr{H}$, $m = 1, 2, \cdots$, *taking* $(x_1, \cdots, x_m)$ *into* $\hat{\hat{\ell}}(\cdot, \cdot; x_1, \cdots, x_m)$ *such that for all regular* (θ, G) *and any* $|\theta_m - \theta| = O(m^{-1/2})$,

$$\int |\hat{\hat{\ell}}_m(x, \theta_m; X_1, \cdots, X_m) - \tilde{\ell}(x, \theta_m, G)|^2 F_{(\theta_m, G)}(dx) \to 0 \tag{3.5}$$

in $P_{(\theta,G)}$ *probability.*

Note that GR(iv) is evidently a necessary condition for adaptive estimation and is the natural generalization of R(iv). Under Condition S*, M-estimates corresponding to a fixed shape are natural candidates for $\tilde{\theta}_n$. In view of Stein's necessary Condition S*, we conjecture that Condition H is necessary for adaptation. W. R. van Zwet pointed out a suggestive inequality bolstering this conjecture (Klaassen, 1980, Theorem 3.2.1). In any case these conditions are sufficient.

THEOREM 3.1. *If Conditions* GR(iv) *and* H *hold, then adaptive estimates exist.*

NOTE. The construction is closely related to that given for adaptive rank tests in the linear model by Hájek (1962). A related construction for Example 1 has been given by Bretagnolle (private communication). See also Hasminskii and Ibragimov (1978).

PROOF. Define $\bar{\theta}_n$ as in (2.3). Let $\{m(n)\}$ be a sequence of subsample sizes with $m(n) = o(n)$. Write m for $m(n)$ and let $\bar{n} = n - m$.
Define

$$\hat{\theta}_n = \bar{\theta}_n + \bar{n}^{-1} \textstyle\sum_{i=m+1}^{n} \hat{\hat{\ell}}(X_i, \bar{\theta}_n; X_1, \cdots, X_m). \tag{3.6}$$

We claim $\{\hat{\theta}_n\}$ is adaptive. By Theorem 6.2,

$$\bar{\theta}_n + \bar{n}^{-1} \textstyle\sum_{i=m+1}^{n} \tilde{\ell}(X_i, \bar{\theta}_n, G)$$

is efficient for every regular (θ, G). Write P_θ for $P_{(\theta,G)}$. Then to prove the theorem it is enough to show

$$\bar{n}^{-1/2} \textstyle\sum_{i=m+1}^{n} \{\hat{\hat{\ell}}_m(X_i, \bar{\theta}_n; X_1, \cdots, X_m) - \tilde{\ell}(X_i, \bar{\theta}_n, G)\} = o_{P_\theta}(1). \tag{3.7}$$

Now we use a trick of Le Cam's and note that we need only establish (3.7) with $\bar{\theta}_n$ replaced by $\theta_n = \theta + t_n n^{-1/2}$, where t_n is an arbitrary convergent deterministic sequence. This follows since $\bar{\theta}_n$ is $\sqrt{n}$ − consistent and the intersection of its range with any sphere of radius $Mn^{-1/2}$ about θ is finite with cardinality bounded independent of n. Having made the replacement, we prove (3.7). Note that R(i) − R(iii) imply that the $\bar{n}$ dimensional product measures of $X_{m+1}, \cdots, X_n$ under P_θ and under P_{θ_n} are contiguous. Therefore, it suffices to prove (3.7) in P_{θ_n} probability. Condition on $X_1, \cdots, X_m$ for this probability. Since $\hat{\hat{\ell}}(\cdot, \cdot; X_1, \cdots, X_m) \in \mathscr{H}$,

$$\int \hat{\hat{\ell}}_m(x, \theta_n; X_1, \cdots, X_m) f(x, \theta_n, G)\mu(dx) = 0 \tag{3.8}$$

and by R(i) − R(iii),

$$\int \tilde{\ell}(x, \theta_n, G) f(x, \theta_n, G)\mu(dx) = 0. \tag{3.9}$$

Therefore

$$\begin{aligned} &E_{\theta_n}[\,|\bar{n}^{-1/2} \textstyle\sum_{i=m+1}^n \{\hat{\tilde{\ell}}_m(X_i, \theta_n; X_1, \cdots, X_m) - \tilde{\ell}(X_i, \theta_n, G)\}|^2 \,|\, X_1, \cdots, X_m] \\ &\quad = \int |\hat{\tilde{\ell}}_m(x, \theta_n; X_1, \cdots, X_m) - \tilde{\ell}(x, \theta_n, G)|^2 f(x, \theta_n, G)\mu(dx) \to 0 \end{aligned} \tag{3.10}$$

in P_θ probability by Condition H and hence, by contiguity again, in P_{θ_n} probability. Claim (3.7) is proved, and the theorem follows. □

NOTES. It is possible to replace Condition H by the following condition H′ which permits separate estimation of $\dot{\ell}$ and I^{-1}.

CONDITION H'. (a) *There exist maps* $\hat{\dot{\ell}}_m (R^k)^m \to \mathscr{H}$ *such that for all regular* (θ, G), $|\theta_m - \theta| = O(m^{-1/2})$

$$\int |\hat{\dot{\ell}}_m(x, \theta_m; X_1, \cdots, X_m) - \dot{\ell}(x, \theta_m, G)|^2 f(x, \theta_m, G)\mu(dx) = o_{P_\theta}(1). \tag{3.11}$$

(b) *There exist estimates* $\hat{I}_m(X_1, \cdots, X_m)$ *of* $I(\theta, G)$ *consistent for all regular* (θ, G).

It is easy to show that if GR(iv) and H′ both hold, and if we define

$$\theta_n^* = \bar{\theta}_n + \bar{n}^{-1} \textstyle\sum_{i=m+1}^n \hat{\dot{\ell}}(X_i, \bar{\theta}_n; X_1, \cdots, X_m)\hat{I}_n^{-} \tag{3.12}$$

then

$$\theta_n^* = \bar{\theta}_n + \bar{n}^{-1} \textstyle\sum_{i=m+1}^n \hat{\dot{\ell}}(X_i, \bar{\theta}_n; X_1, \cdots, X_m) I^{-1}(\theta, G) + o_{P_\theta}(n^{-1/2}) \tag{3.13}$$

and θ_n^* is adaptive.

A natural choice of $\hat{I}_n$ is provided by

$$\hat{I}_n = \bar{n}^{-1} \textstyle\sum_{i=m+1}^n \hat{\dot{\ell}}^T \hat{\dot{\ell}}(X_i, \theta_n; X_1, \cdots, X_m) \tag{3.14}$$

We show in Section 6.2 that this choice of $\hat{I}_n$ is consistent for regular (θ, G) provided that GR(iv) and (3.11) hold, and if

$$m^{-1} \textstyle\sum_{i=1}^m \dot{\ell}^T \dot{\ell}(X_i, \theta_m, G) \to I(\theta, G) \tag{3.15}$$

in P_θ probability for all regular (θ, G).

These are the results we will apply to Example 2 and which are applicable to other situations where all of θ is estimable. To deal with Examples 3 and 4 we need an extension of our theory. First we study the analogue of Condition S* when we only ask that $q(\theta)$, rather than all of θ, be estimated adaptively. Stein considers this question in a slightly different formulation. He writes $\theta = (q, t)$ with $q = q(\theta)$ and t, the rest of θ, is a nuisance parameter, and he introduces the model $\{F_{(\theta, G_\eta)}\}$. He notes that adaptive estimation of q is possible only if the upper left-hand corner of the inverse of the information matrix for (q, t) with $\eta = \eta_0$ fixed is the same as the upper left-hand corner of the inverse of the information matrix for (q, t, η) evaluated at η_0. We do not pursue further his matrix formulation of this condition, but only note that in the presence of convexity Condition C, Stein's condition is heuristically equivalent to the d equations

CONDITION S* (*generalized*).

$$\int \dot{\ell}(x, \theta, G_0) I^{-1}(\theta, G_0) \dot{q}^T(\theta) f(x, \theta, G_1)\mu(dx) = 0$$

for every shape $G_0, G_1 \in \mathscr{G}$. For $q(\theta) = \theta$, $\dot{q}$ is the identity and our more general formulation of S* agrees with our old one.

New difficulties are introduced by the possible lack of identifiability of θ. Of course we need to have q identifiable. That is, if

$$F_{(\theta_0, G_0)} = F_{(\theta_1, G_1)} = F \tag{3.16}$$

then

$$q(\theta_0) = q(\theta_1).$$

But adaptation requires more. If F can be embedded in both $\mathscr{F}_{G_0}$ and $\mathscr{F}_{G_1}$ as in (3.16), then the information bound for estimation of q must be the same in both parametric families. That is, (3.16) implies

$$\dot{q}(\theta_0) I^-(\theta_0, G_0) \dot{q}^T(\theta_0) = \dot{q}(\theta_1) I^-(\theta_1, G_1) \dot{q}^T(\theta_1). \tag{3.17}$$

This condition is satisfied in all our examples because if $\mathscr{F}_{G_0}$ and $\mathscr{F}_{G_1}$ have a member in common then they are the same, or, rather, one is a smooth relabelling of the other. For instance, in Example 3, (3.16) holds if and only if G_1 is obtained from G_0 by a translation. We shall use this structural feature in a stronger way to reduce $\mathscr{G}$ and make θ identifiable. Here is a formal statement of our structural assumptions. They are obviously satisfied in Examples 3 and 4.

Assumption A1. *Either $\mathscr{F}_{G_0} = \mathscr{F}_{G_1}$ or $\mathscr{F}_{G_0} \cap \mathscr{F}_{G_1} = \varnothing$, for all $G_0, G_1 \in \mathscr{G}$.*

Assumption A2. *There exists $T \subset R^{p-d}$ and a smoothly invertible map from Θ to $Q \times T$ where $Q = q(\Theta)$ which carries θ into $(q(\theta), t(\theta))$. That is, we can identify q with a piece of θ.*

Assumption A3. *If we replace θ by (q, t) and $\mathscr{F}_{G_0} = \mathscr{F}_{G_1}$ there exists a unique smoothly invertible mapping $\tau(q, \cdot)$ of T into itself defined by $F_{(q,t,G_0)} = F_{(q,\tau,G_1)}$.*

Assumption A1 implies that there exists an "identifying subset" $\mathscr{G}_0 \subset \mathscr{G}$ such that (i) $\mathscr{F} = \{F_{(\theta,G)} : G \in \mathscr{G}_0, \theta \in \Theta\}$, and (ii) θ is identifiable when G is restricted to $\mathscr{G}_0$ provided that it is identifiable in each $\mathscr{F}_G$. We can select $\mathscr{G}_0$ as a set of representatives of the equivalence classes generated by the relation $G_1 \equiv G_2 \Leftrightarrow \mathscr{F}_{G_1} = \mathscr{F}_{G_2}$. For instance, in Example 3 we can take $\mathscr{G}_0 = \{G : \mu(G) = 0\}$ where μ is a location parameter. As we noted, Assumptions A2 and A3 imply that if (a) $\mathscr{F}_G = \mathscr{F}_{G_0}$, $G_0 \in \mathscr{G}_0$, and (b) $F_{(\theta,G)} = F_{(\theta_0,G_0)}$, then $q(\theta) = q(\theta_0)$ and (3.16) holds. That is, it does not matter in which parametric model $\mathscr{F}_G$ we embed a distribution F. The value of q and the ease with which q can be estimated remain the same. Since we can talk about estimation of θ for $(\theta, G) \in \Theta \times \mathscr{G}_0$ it is natural to propose the following extensions of the conditions for $\sqrt{n}$-consistency and appropriate consistent estimation of score functions.

Generalized Condition GR(iv). *There exists $\mathscr{G}_0$ satisfying* (i) *and* (ii) *above and estimates $\{\tilde{\theta}_n\}$ such that*

$$n^{1/2}(\tilde{\theta}_n - \theta) = O_{P_{(\theta,G)}}(1)$$

for all (θ, G), $G \in \mathscr{G}_0$.

We now redefine $\tilde{\ell}$, $\mathscr{H}$ *for given* q. Our definitions agree with the old ones when q is the identity. Let

$$\mathscr{H} = \left\{ h : h \text{ maps } R^k \times \Theta \text{ into } R^d \text{ so that} \right.$$
$$\left. \int h(x, \theta) f(x, \theta, G) \mu(dx) = 0 \quad \text{for all} \quad (\theta, G) \right\}.$$

$$\tilde{\ell}(x, \theta, G) = \ell(x, \theta, G) I^-(\theta, G) \dot{q}^T(\theta). \tag{3.18}$$

Condition H is now generalized as was condition GR(iv), merely by substituting $\mathscr{G}_0$ for $\mathscr{G}$. The easy extension of Theorem 3.1 is as follows.

THEOREM 3.2. *If Assumptions* A1–A3 *and the generalized conditions* GR(iv) *and* H *hold, then adaptive estimates* $\{\hat{q}_n\}$ *of* $q(\theta)$ *exist.*

The proof is the same as for Theorem 3.1 when we propose as estimate

$$\hat{q}_n = q(\bar{\theta}_n) + \bar{n}^{-1} \textstyle\sum_{i=m+1}^n \hat{\bar{\ell}}_m(X_i, \bar{\theta}_n; X_1, \cdots, X_m). \tag{3.19}$$

4. Adaptation in Examples 1–4. For the examples we leave verification of the trivial structural Assumptions A1 through A3 to the reader. In each example we shall proceed through the following steps:

Step A. Formally verify Stein's orthogonality Condition S* and in the process construct what we can think of as the "space of possible score functions" $\mathscr{H}$ or a suitable subset $\mathscr{H}_0$.

Step B. Find a suitable identifying subset $\mathscr{B}_0$ and construct $\sqrt{n}$-consistent estimates $\{\tilde{\theta}_n\}$ so as to satisfy GR(iv).

Step C. Construct score function estimates $\hat{\bar{\ell}}$ satisfying (3.5) and taking values in $\mathscr{H}_0$ i.e. satisfy Condition H for the appropriate consistent estimation of score functions, or satisfy its modification H′ providing for separate estimation of $\dot{\ell}$ and I.

Since Example 1 is a special case of Example 2 and has already been dealt with satisfactorily, we begin with Example 2. For convenience from now on we write P for P_θ.

EXAMPLE 2. *Step A.* If the distribution of C has density r with respect to some ν, and if G has density g, then $X = (C, Y)$ has density (with respect to the product measure)

$$f(c, y, \theta, G) = r(c)g(y - c\theta^T), \tag{4.1}$$

and

$$\dot{\ell}(c, y, \theta, G) = c\frac{g'}{g}(y - c\theta^T). \tag{4.2}$$

Then

$$E_{(\theta,G_0)}\dot{\ell}(C, Y, \theta, G) = E_{(\theta,G_0)}\left\{C\frac{g'(\varepsilon)}{g(\varepsilon)}\right\} = E(C)E_{G_0}\left\{\frac{g'(\varepsilon)}{g(\varepsilon)}\right\} = 0,$$

since g'/g is antisymmetric and G_0 is symmetric about 0. Thus, Condition S* is satisfied and by our argument, $\mathscr{H} \supset \mathscr{H}_0$ where $h \in \mathscr{H}_0$ if and only if

$$h(c, y, \theta) = c\psi(y - c\theta^T) \tag{4.3}$$

for ψ bounded and antisymmetric, i.e.

$$\psi(y) = -\psi(-y). \tag{4.4}$$

So we will use score function estimates of the form (4.3).

Step B. Let $\psi: R \to R$ be such that ψ is twice continuously differentiable, with ψ and its derivatives bounded. Suppose, moreover, that $\psi' > 0$ and that ψ is antisymmetric. Let $\{\tilde{\theta}_n\}$ be the M-estimates corresponding to ψ, i.e., the unique solutions of

$$\textstyle\sum_{i=1}^n C_{ij}\psi(Y_i - C_i\tilde{\theta}_n^T) = 0, \quad j = 1, \cdots, p, \tag{4.5}$$

where $X_i = (C_i, Y_i)$, $C_i = (C_{i1}, \cdots, C_{ip})$. Then by Huber's theorem (Huber, 1973), $\{\tilde{\theta}_n\}$ are $\sqrt{n}$-consistent. (This is just the construction suggested in the previous section.)

Step C. By modifying the arguments of Hájek (1972) it is easy to see that (θ, G) is regular if g is absolutely continuous with derivative g' and if $I(G)$, the Fisher information for location given in Section 2, is finite. The converse is also true (proof available from author).

By (4.2) we calculate

$$\tilde{\dot{\ell}}(c, y, \theta, G) = c\frac{g'}{g}(y - c\theta^T)\{E(C^TC)I(G)\}^{-1}, \tag{4.6}$$

where the last term is just $I^{-1}(\theta, G)$. To apply Condition H or H′ we need to estimate g'/g and $I(G)$. This is achieved by the following lemma whose proof is given in Section 6.1.

LEMMA 4.1. *Let* $\varepsilon_1, \varepsilon_2, \cdots$ *be* i.i.d. *random variables. There exists a sequence of function estimates* $q_m: R \times R^m \to R$, $m = 1, 2, \cdots$, *such that* q_m *is bounded for each* m *and such that as* $m \to \infty$

$$\int \left\{q_m(y; \varepsilon_1, \cdots, \varepsilon_m) - \frac{g'(y)}{g(y)}\right\}^2 g(y)\, dy \to 0 \tag{4.7}$$

in probability whenever the common d.f. *of the* ε_i *is* G *with density* g *and* $I(G) < \infty$.

We proceed to show how to estimate $\dot{\ell}$ and $I(G)$ separately and verify Condition H′. Let

$$\hat{\varepsilon}_i = Y_i - C_i\bar{\theta}_m^T(X_1, \cdots, X_m), \quad i = 1, \cdots, m, \tag{4.8}$$

be the residuals with respect to the "discretized" estimate based on the first m observations. Define

$$\psi_m(y; X_1, \cdots, X_m) = \tfrac{1}{2}\{q_m(y; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m) - q_m(-y; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m)\} \tag{4.9}$$

and

$$\hat{\dot{\ell}}_m(c, y, \theta; X_1, \cdots, X_m) = c\psi_m(y - c\theta^T; X_1, \cdots, X_m). \tag{4.10}$$

Clearly $\hat{\dot{\ell}}(\cdot; X_1, \cdots, X_m) \in \mathscr{H}_0$ and

$$\begin{aligned} &\int |\hat{\dot{\ell}}_m(c, y, \theta_m; X_1, \cdots, X_m) - \dot{\ell}(c, y, \theta_m, G)|^2 f(c, y, \theta_m, G)\, dy\, \nu(dc) \\ &= \int c\left|\psi_m(y - c\theta_m^T; X_1, \cdots, X_m) - \frac{g'}{g}(y - c\theta_m^T)\right|^2 c^T g(y - c\theta_m^T)\, dy\, \nu(dc) \qquad (4.11) \\ &\le \left[\left\{\int \left|q_m(y; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m) - \frac{g'}{g}(y)\right|^2 g(y)\, dy\right\} ECC^T\right]. \end{aligned}$$

Now let $\theta_m = \theta + t_m$, where t_m and $c_1, \cdots, c_m$ are p-dimensional vectors such that $|t_m| = O(m^{-1/2})$ and $\sum_{i=1}^m c_i t_m^T t_m c_i^T$ is bounded independent of m. Then the sequence of m-dimensional product measures induced by $\varepsilon_1, \cdots, \varepsilon_m$ and $\varepsilon_1 - c_1 t_m^T, \cdots, \varepsilon_m - c_m t_m^T$ are contiguous if $I(G) < \infty$ (Hájek and Sidák, 1967, page 211). Since ECC^T is finite, if $|t_m| = O(m^{-1/2})$, $\sum_{i=1}^m c_i t_m^T t_m c_i^T = O_{P_\theta}(1)$. Thus, by Lemma 4.1,

$$\int \left|q_m(y; \varepsilon_1 - C_1 t_m^T, \cdots, \varepsilon_m - C_m t_m^T) - \frac{g'}{g}(y)\right|^2 g(y)\, dy \to_{P_\theta} 0. \tag{4.12}$$

But, as usual, by the structure of $\bar{\theta}_m$ and its $m^{1/2}$-consistency, this result is enough to establish

$$\int \left\{q_m(y; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m) - \frac{g'}{g}(y)\right\}^2 g(y)\, dy \to_{P_\theta} 0. \tag{4.13}$$

Substituting in (4.11), we see that $\hat{\dot{\ell}}_m$ is a consistent estimate of $\dot{\ell}$ in the sense of part (a) of Condition H′, in (3.11).

There are various ways to construct $\hat{I}_n$. For instance, we can verify (3.15) in this case as follows:

$$m^{-1}\sum_{i=1}^m \dot{\ell}^T\dot{\ell}(X_i, \theta_m, G) = m^{-1}\sum_{i=1}^m C_i^T C_i\left(\frac{g'}{g}\right)^2(Y_i - C_i\theta_m^T)$$

(4.14)

$$\to_{P_{\theta_m}} E(C^T C)I(G) = I(\theta, G)$$

by the weak law of large numbers. By contiguity we can replace θ_m by θ in P_{θ_m}. This yields as the consistent estimate of (3.14),

$$\hat{I}_n^{(1)} = \bar{n}^{-1}\sum_{i=m+1}^n C_i^T C_i\psi_m^2(Y_i - C_i\bar{\theta}_n^T; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m). \tag{4.15}$$

A more familiar alternative, which may similarly be shown to work, is

$$\hat{I}_n^{(2)} = (n^{-1}\sum_{i=1}^n C_i^T C_i)\bar{n}^{-1}\sum_{i=m+1}^n \psi_m^2(Y_i - C_i\bar{\theta}_m^T; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m). \tag{4.16}$$

We have proved the following result.

THEOREM 4.1. *Let $\tilde{\theta}_n$ be defined as in* (4.5), *ψ_m as in* (4.9). *Let*

$$\hat{\theta}_n = \bar{\theta}_n + \bar{n}^{-1}\sum_{i=m+1}^n C_i\psi_m(Y_i - C_i\bar{\theta}_m^T; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m) \tag{4.17}$$

where $\hat{I}_n$ is given by (4.15) *or* (4.16). *Then $\{\hat{\theta}_n\}$ is adaptive in Example* 2.

EXAMPLE 3.

Step A. If $c = (c^\circ, 1)$, $q(\theta) = (\theta_1, \cdots, \theta_{p-1})$ and $\tilde{\ell}$ is defined by (3.18), we get

$$\tilde{\ell}(c, y, \theta, G) = (c^\circ - EC^\circ)(\text{Var } C^\circ)^{-1}\frac{g'}{g}(y - c\theta^T)I^{-1}(G). \tag{4.18}$$

Thus, formally

$$E_{(\theta, G_0)}\tilde{\ell}(X, \theta, G) = E(C^\circ - EC^\circ)(\text{Var } C^\circ)^{-1}E\frac{g'}{g}(\varepsilon)I^{-1}(G) = 0$$

and Condition S* is satisfied. In view of (4.18) it is natural to choose

$$\mathscr{H}_0 = \{h: h(c, y, \theta) = (c^\circ - EC^\circ)(\text{Var } C^\circ)^{-1}\psi(y - c\theta^T), \psi \text{ bounded}\}. \tag{4.19}$$

Step B. Let ψ be as in Step B of Example 2 and define

$$\mathscr{G}_0 = \left\{G: \int \psi(y)G(dy) = 0\right\}. \tag{4.20}$$

Evidently $\mathscr{G}_0$ is an identifying subset and, by Huber's theorem, $\{\tilde{\theta}_n\}$ corresponding to ψ are $\sqrt{n}$-consistent when G is restricted to $\mathscr{G}_0$.

Step C. A possible definition of $\hat{\tilde{\ell}}$ is just

$$\hat{\tilde{\ell}}_m(c, y, \theta; X_1, \cdots, X_m) = (c^\circ - EC^\circ)(\text{Var } C^\circ)^{-1}q_m(y - c\theta^T; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m)\hat{I}^{-1}, \tag{4.21}$$

where

$$\hat{I} = \bar{n}^{-1}\sum_{i=m+1}^n q_m^2(Y_i - C_i\bar{\theta}_n^T; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m), \tag{4.22}$$

q_m is given in Lemma 4.1 and the $\hat{\varepsilon}_i$ are defined by (4.8). That $\hat{\tilde{\ell}}$ works is evident by the same argument as we gave for Theorem 4.1, since regular (θ, G) again correspond to $I(G) < \infty$. This is not satisfactory, however, because the resultant estimates depend on the first and second moments of the unknown distribution of C°. We claim that estimating these does just as well. Here is one way of proceeding. Define

$$\bar{C}_n^\circ = n^{-1} \textstyle\sum_{i=1}^n C_i^\circ \tag{4.23}$$

$$\hat{\text{V}}\text{ar } C^\circ = n^{-1} \textstyle\sum_{i=1}^n (C_i^\circ - \bar{C}_n^\circ)^T(C_i^\circ - \bar{C}_n^\circ).$$

Let

$$\hat{q}_n = \bar{\theta}_n^{(p-1)} + \bar{n}^{-1} \textstyle\sum_{i=m+1}^n (C_i^\circ - \bar{C}_n^\circ)(\hat{\text{V}}\text{ar } C_n^\circ)^- q_m(Y_i - C_i\bar{\theta}_n^T; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m)\hat{I}^{-1} \tag{4.24}$$

where $\bar{\theta}_n^{(p-1)}$ is the vector of the initial $p-1$ elements of $\bar{\theta}_n$.

THEOREM 4.2. *The estimates $\hat{q}_n$ defined by* (4.24) *adaptively estimate* $(\theta_1, \cdots, \theta_{p-1})$ *in Example* 3.

PROOF. We know that

$$\bar{n}^{-1} \textstyle\sum_{i=m+1}^n (C_i^\circ - EC^\circ)q_m(Y_i - C_i\bar{\theta}_n^T; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m)(\text{Var } C^\circ)^- = o_P(n^{-1/2}) \tag{4.25}$$

and

$$\hat{\text{V}}\text{ar } C^\circ = \text{Var } C^\circ + o_P(1). \tag{4.26}$$

Therefore, replacing Var C° by $\hat{\text{V}}$ar C° in (4.21) will still lead to adaptive estimates. Thus to establish that the estimates given by (4.24) are adaptive it suffices to prove that

$$\bar{n}^{-1} \textstyle\sum_{i=m+1}^n (\bar{C}_n^\circ - EC^\circ)(\hat{\text{V}}\text{ar } C^\circ)^- q_m(Y_i - C_i\bar{\theta}_n^T; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m) = o_P(n^{-1/2}) \tag{4.27}$$

or, since $\bar{C}_n^\circ - EC^\circ = O_P(n^{-1/2})$, that

$$\bar{n}^{-1} \textstyle\sum_{i=m+1}^n q_m(Y_i - C_i\bar{\theta}_n^T; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m) = o_P(1). \tag{4.28}$$

To prove (4.28) we show that we can replace q_m by g'/g and $Y_i - c_i\bar{\theta}_n^T$ by ε_i and then apply the law of large numbers. Details are given in Section 6.2. □

EXAMPLE 4. *Step A*. In this case if $\theta = (\mu, [V])$, then

$$f(x, \theta, G) = \{\det(V)\}^{1/2}\gamma(\{(x-\mu)V(x-\mu)^T\}^{1/2}) \tag{4.29}$$

where det denotes determinant, and γ maps R^+ into itself. Of course, $\gamma(|x|)$ is the density of G. We want to estimate

$$q(\mu, [V]) = (\mu, q_0([V])) \tag{4.30}$$

where q_0 is any homogeneous function of $[V]$. A "most general" choice is $q_0([V]) = [V]/\text{tr}(V)$. We can write, for (θ, G_0) regular,

$$\dot{\ell}(x, \mu, G_0)I^{-1}(\theta, G_0) = (\psi^\circ(x, \mu, V), [\chi^\circ(x, \mu, V)])$$

where ψ° is $1 \times k$, χ° is $k \times k$ symmetric, and $[\chi]$ denotes the $k(k+1)/2$ dimensional vector of the lower half of χ. It is shown in Section 6.3 that

$$\psi^\circ(x, \mu, V) = \psi^\circ((x-\mu)V^{1/2}, 0, J)V^{-1/2} \tag{4.31}$$

$$\chi^\circ(x, \mu, V) = V^{1/2}\chi^\circ((x-\mu)V^{1/2}, 0, J)V^{1/2}, \tag{4.32}$$

where J is the $k \times k$ identity matrix. We further show in Section 6.3 that, if $|\cdot|$ is the Euclidean norm and $\gamma_0(|x|)$ is the density of G_0, then

$$\psi^\circ(x, 0, J) = -\frac{x}{|x|}\frac{\gamma_0'}{\gamma_0}(|x|)kI_1^{-1}(G_0) \tag{4.33}$$

and

$$\chi^{\circ}_{ij}(x, 0, J) = \begin{cases} I_2^{-1}(G_0)k(k+2)\dfrac{x_i x_j}{|x|}\dfrac{\gamma_0'}{\gamma_0}(|x|), & i \neq j, \\ 2\left\{I_2(G_0)\dfrac{3}{k(k+2)} - 1\right\}^{-1}\left\{\dfrac{x_i^2}{|x|}\dfrac{\gamma_0'}{\gamma_0}(|x|) + 1\right\}, & i = j, \end{cases} \tag{4.34}$$

where

$$I_1(G) = c_k \int_0^{\infty} r^{k-1} \frac{[\gamma']^2}{\gamma}(r)\, dr \tag{4.35}$$

$$I_2(G) = c_k \int_0^{\infty} r^{k+1} \frac{[\gamma']^2}{\gamma}(r)\, dr \tag{4.36}$$

and c_k is the surface area of the unit sphere in R^k. Then by (4.31) and (4.32),

$$\begin{aligned} &E_{(\theta,G)}\{\psi^{\circ}(X, \mu, V), [\chi^{\circ}(X, \mu, V)]\}\dot{q}^T(\theta) \\ &\qquad = E_{(0,[J],G)}\{\psi^{\circ}(X, 0, J)V^{-1/2}, [V^{1/2}\chi^{\circ}(X, 0, J)V^{1/2}]\}\dot{q}^T(\theta). \end{aligned} \tag{4.37}$$

Moreover, if $i \neq j$, X°_{ij} changes sign if all the coordinates of x other than x_i are left unchanged while $x_i \to -x_i$. Since if $\theta = (0, [J])$, all the X_i are identically distributed and the distributions of $(X_1, \cdots, X_k)$ and $(\pm X_1, \cdots, \pm X_k)$ are the same, we conclude that

$$E_{(0,[J],G)}\psi^{\circ}(X, 0, J) = 0 \tag{4.38}$$

$$E_{(0,[J],G)}\chi^{\circ}(X, 0, J) = cJ, \tag{4.39}$$

where c depends on G and G_0. Therefore

$$E_{(0,[J],G)}[V^{1/2}\chi^{\circ}(X, 0, J)V^{1/2}] = c[V]. \tag{4.40}$$

Substituting (4.38) and (4.40) back into (4.37) we find that all components of (4.37) vanish either by (4.38) or by Euler's equation $\sum_{k \geq \ell} v_{k\ell}\, \partial q_0/\partial v_{k\ell} = 0$.

The orthogonality Condition S* follows and our argument makes it clear that $\mathscr{H}$ defined in (3.3), contains the set $\mathscr{H}_0$ of $h(x, \theta)$ defined by

$$h(x, \theta) = (\psi((x-\mu)V^{1/2})V^{-1/2}, [V^{1/2}\chi((x-\mu)V^{1/2})V^{1/2}])\dot{q}^T(\theta), \tag{4.41}$$

where ψ is $1 \times k$ and χ is symmetric $k \times k$ with forms

$$\psi(x) = \omega(|x|)\frac{x}{|x|}a_1 \tag{4.42}$$

$$\chi_{ij}(x) = \begin{cases} \omega(|x|)\dfrac{x_i x_j}{|x|}a_2, & i \neq j, \\ \left\{\omega(|x|)\dfrac{x_i^2}{|x|} + 1\right\}a_3, & i = j, \end{cases} \tag{4.43}$$

where ω is bounded and a_1, a_2, a_3 are constant. Clearly $\mathscr{H}$ is much bigger than $\mathscr{H}_0$, but $\mathscr{H}_0$ is the space of natural estimates of ℓ.

Step B. Thanks to Maronna (1976) we can find an identifying subset $\mathscr{G}_0$ and corresponding $\sqrt{n}$ – consistent $\tilde{\theta}_n$ as follows. Let u_1 and u_2 be functions on R^+. Define the M-estimate $(\tilde{\mu}_n, \tilde{V}_n)$ corresponding to u_1 and u_2 to be any solution of

$$\begin{aligned} &n^{-1}\textstyle\sum_{i=1}^n u_1(\{(X_i - \tilde{\mu}_n)\tilde{V}_n(X_i - \tilde{\mu}_n)^T\}^{1/2}) = 0 \\ &n^{-1}\textstyle\sum_{i=1}^n u_2(\{(X_i - \tilde{\mu}_n)\tilde{V}_n(X_i - \tilde{\mu}_n)^T\})(X_i - \tilde{\mu}_n)^T(X_i - \tilde{\mu}_n) = [\tilde{V}_n]^{-1} \end{aligned} \tag{4.44}$$

if one exists, and arbitrarily otherwise.

It is easy to see that the maximum likelihood estimates for a particular G are of this type. Let u_1, u_2 satisfy conditions (A) – (D) on page 53 of Maronna (1976). In addition, if $\psi_i(s) = su_i(s)$, $i = 1, 2$, suppose that $s\psi_j'(s)$ are bounded, $j = 1, 2$, and $\psi_1' > 0$. By Theorem 5.6 of Maronna, under these conditions $n^{1/2}(\tilde{\mu}_n - \tilde{\mu}, \tilde{V}_n - \tilde{V}) = O_P(1)$ for all $F \in \mathscr{F}$ where $\tilde{\mu}(\mu, V, G)$, $\tilde{V}(\mu, V, G)$ satisfy *uniquely*

$$\int u_1(\{(x-\tilde{\mu})\tilde{V}(x-\tilde{\mu})^T\}^{1/2})(x-\tilde{\mu})f(x, \theta, G)\,dx = 0 \tag{4.45}$$

$$\int u_2(\{(x-\tilde{\mu})\tilde{V}^T(x-\tilde{\mu})^T\})(x-\tilde{\mu})^T(x-\tilde{\mu})f(x, \theta, G)\,dx = [\tilde{V}]^{-1}. \tag{4.46}$$

It is clear by the unicity of $\tilde{\mu}$, $\tilde{V}$ that

$$\tilde{\mu}(\mu, V, G) = \mu, \tag{4.47}$$

$$\tilde{V}(\mu, V, G) = c(G)V, \tag{4.48}$$

where $c(G)$ is that measure of scale which is the unique solution of the equation

$$E\{u_2(c\,\varepsilon\,\varepsilon^T)\} = \frac{1}{c};$$

existence is guaranteed by the monotonicity of u_2. Clearly we can take as an identifying subset

$$\mathscr{G}_0 = \{G : c(G) = 1\} \tag{4.49}$$

and $\tilde{\theta}_n = (\tilde{\mu}_n, \tilde{V}_n)$ defined by (4.44).

Step C. It may be shown that regularity of (θ, G) is equivalent to absolute continuity of γ on $(0, \infty)$ and finiteness of $I_1(G)$ and $I_2(G)$. (Proof available from author.) We will show how to construct adaptive estimates of $q_0(V)$ in a simple fashion and then discuss the simultaneous adaptive estimation of μ.

Note that if X has density given by (4.29), then $\log|(X-\mu)V^{1/2}|$ has density j given by

$$j(z) = c_k e^{kz}\gamma(e^z). \tag{4.50}$$

Thus

$$\frac{\gamma'}{\gamma}(y) = y^{-1}\left\{\frac{j'}{j}(\log y) - k\right\}, \quad y > 0, \tag{4.51}$$

and this leads to the following construction of an estimate of γ'/γ.

Let $\bar{\mu}_m$ be obtained by discretizing $\tilde{\mu}_m$ as usual while $[\tilde{V}_m]$ is the closest member of the $m^{-1/2}$ lattice to $\tilde{V}_m$ which itself corresponds to a positive definite matrix. Let

$$z_{im} = \log|(X_i - \bar{\mu}_m)\bar{V}_m^{1/2}|, \quad i = 1, \cdots, m,$$

and define

$$\omega_m(y; X_1, \cdots, X_m) = y^{-1}\{\acute{q}_m(\log y; z_{1m}, \cdots, z_{mm}) - k\}. \tag{4.52}$$

We claim that

$$\int |x|^2\left|\omega_m(|x|; X_1, \cdots, X_m) - \frac{\gamma'}{\gamma}(|x|)\right|^2\gamma(|x|)\,dx \to 0 \tag{4.53}$$

in P_θ probability if (θ, G) is regular. The proof follows the usual lines. By construction of $\bar{\mu}_m$, $\bar{V}_m$ it is possible to treat them as deterministic sequences such that $|\bar{\mu}_m - \mu|$ and $|\bar{V}_m - V| = O(m^{-1/2})$. Since (θ, G) is regular the m-dimensional product measures induced by $\varepsilon_1, \cdots, \varepsilon_m$ and $(X_1 - \bar{\mu}_m)\bar{V}_m^{1/2}, \cdots, (X_m - \bar{\mu}_m)\bar{V}_m^{1/2}$ are contiguous. If we also use (4.51) we can conclude that (4.53) is equivalent to

$$\int\left|q_m(\log|x|; \log|\varepsilon_1|, \cdots, \log|\varepsilon_m|) - \frac{j'}{j}(\log|x|)\right|^2\gamma(|x|)\,dx \to 0 \tag{4.54}$$

in probability whenever $\varepsilon_1, \cdots, \varepsilon_m$, are i.i.d. with common distribution G such that $I_1(G)$ and $I_2(G)$ are finite. But the integral in (4.54) equals

$$\int_{-\infty}^{\infty} \left| q_m(z; \log|\varepsilon_1|, \cdots, \log|\varepsilon_m|) - \frac{j'}{j}(z) \right|^2 g(z)\, dz. \tag{4.55}$$

Moreover,

$$\int_{-\infty}^{\infty} \frac{(j')^2}{j}(z)\, dz = \int_{-\infty}^{\infty} \left\{ e^z \frac{\gamma'}{\gamma}(e^z) + k \right\}^2 g(z)\, dz = I_2(G) - k^2 \tag{4.56}$$

using integration by parts. Thus the integral in (4.55) tends to 0 whenever $I_2(G) < \infty$ by Lemma 4.1 and (4.54) and hence (4.53) holds. Now that we have an estimate $\omega_m(\cdot; X_1, \cdots, X_m)$ of γ'/γ we can estimate $I_2(G)$ by, for instance, splitting our preliminary sample of m, taking $m = 2\ell$ and letting

$$\hat{I}_2 = \ell^{-1} \textstyle\sum_{i=\ell+1}^{m} q_m^2(z_{im}; z_{1m}, \cdots, z_{\ell m}) + k^2. \tag{4.57}$$

Evidently $\hat{I}_2$ depends only on $X_1, \cdots, X_m$. Moreover, we can argue as for (4.28) that, whenever (θ, G) is regular,

$$\hat{I}_2 \to I_2(G) \text{ in probability.} \tag{4.58}$$

Now define $\hat{\chi}_0(\cdot, O, J)$ by substituting $\hat{I}_2$ for $I_2(G_0)$ and $\omega_m(\cdot; X_1, \cdots, X_m)$ for γ_0'/γ_0 in (4.34) and let

$$\hat{\hat{\ell}}_m(x, \theta; X_1, \cdots, X_m) = [V_m^{1/2} \hat{\chi}_0((x-\mu)V^{1/2}, O, J) V_m^{1/2}] \dot{q}_0^T([V]). \tag{4.59}$$

This is the natural estimate of $\tilde{\ell}$ corresponding to $q_0([V])$. Now after some algebra, if $\theta_m = (\mu_m, [V_m])$,

$$\begin{aligned} &\int |\hat{\hat{\ell}}_m(x, \theta_m; X_1, \cdots, X_m) - \dot{\ell}(x, \theta_m, G) I^{-1}(\theta_m, G)(0, \dot{q}_0([V]))^T|^2 f(x, \theta_m, G)\, dx \\ &= O_P\Bigg(\int |(x-\mu_m)V_m^{1/2}|^2 \Bigg| \omega_m((x-\mu_m)V_m^{1/2}; X_1, \cdots, X_m) \\ &\qquad - \frac{\gamma'}{\gamma}(|(x-\mu_m)V_m^{1/2}|) \Bigg|^2 f(x, \theta_m, G)\, dx \Bigg) + O_P(\hat{I}_2 - I_2). \end{aligned} \tag{4.60}$$

But the right-hand side of (4.60) is $o_p(1)$ by (4.53) and (4.58). From (4.60) and the structure of $\hat{\hat{\ell}}$ we see that $\hat{\hat{\ell}}$ falls in $\mathscr{H}_0$ given by (4.41) and is appropriately consistent. We have proved the following result.

THEOREM 4.3. *In Example* 4, *if we define*

$$\hat{q}_{on} = q_0([\bar{V}_n]) + \bar{n}^{-1} \textstyle\sum_{i=m+1}^{n} \hat{\hat{\ell}}_m(X_i, \bar{\theta}_n; X_1, \cdots, X_m), \tag{4.61}$$

then $\{\hat{q}_{on}\}$ *is an adaptive estimate of* $q_0([V])$.

To estimate μ simultaneously and adaptively using the estimate of γ'/γ we need to show that

$$\int \left| \omega_m(|x|; X_1, \cdots, X_m) - \frac{\gamma'}{\gamma}(|x|) \right|^2 \gamma(|x|)\, dx \to 0 \tag{4.62}$$

in probability, or equivalently that

$$\int_{-\infty}^{\infty} e^{-2z} \left| q_m(z; \log|\varepsilon_1|, \cdots, \log|\varepsilon_m|) - \frac{j'}{j}(z) \right|^2 g(z)\, dz \tag{4.63}$$

in probability. Unfortunately, to show (4.63) we need

$$\int_{-\infty}^{\infty} e^{-2z}\frac{(j')^2}{j}(z)\,dz = c_k\int_0^{\infty} y^{k-1}\left\{\frac{\gamma'}{\gamma}(y) + ky^{-1}\right\}^2\gamma(y)\,dy < \infty \tag{4.64}$$

and this happens if $I_1(G) < \infty$ and

$$\int_0^{\infty} y^{k-3}\gamma(y)\,dy < \infty, \tag{4.65}$$

a superfluous condition.

To get rid of (4.65) we need to estimate γ'/γ differently by smoothing the multivariate empirical distribution of $(X_i - \bar{\mu}_m)\bar{V}_m^{1/2}$ and constructing an estimate of γ'/γ out of the first partial derivatives of the smoothed empirical distribution. This can be done but we omit the tedious and rather technical definition of the estimate and the necessary argument.

5. Questions raised by this work and other issues in adaptive estimation.

5.1 *When is adaptation not possible*? We have seen heuristically the necessity of the $\sqrt{n}$-consistency condition GR(iv) and the orthogonality Condition S when there are no nuisance parameters. In parametric models $\sqrt{n}$-consistency is available under mild smoothness and identifiability conditions while orthogonality is special. Orthogonality seems special in these nonparametric nuisance parameter models as well. We illustrate with a famous example of Neyman and Scott. The failure of adaptation in this case was already noted by Wolfowitz (1953).

EXAMPLE 5. *Estimation in Model II.* Suppose $X_i = (X_{i1}, X_{i2})$, $i = 1, \cdots, n$, such that

$$X_{ij} = \mu_i + \varepsilon_{ij}, \quad j = 1, 2, \tag{5.1}$$

where the ε_{ij} are independent identically distributed $\mathcal{N}(0, \theta)$, and the μ_i are independent and identically distributed with common distribution G. Let $\Theta = R^+$, $\mathcal{G} = \{$all distributions on $R\}$. It is easy to see that all (θ, G) are regular, and there is a natural $\sqrt{n}$-consistent estimate, the best unbiased estimate when the μ_i are treated as constants,

$$\bar{\theta}_n = \frac{1}{2n}\sum_{i=1}^n (X_{i1} - X_{i2})^2. \tag{5.2}$$

Thus Condition GR(iv) holds. But Condition H does not. For instance, take G_0 to be point mass at 0. Then

$$\dot{\ell}(x_1, x_2, \theta, G_0) = \frac{1}{\theta}\left\{\frac{(x_1^2 + x_2^2)}{2\theta} - 1\right\} \tag{5.3}$$

and

$$E_{(\theta,G)}\dot{\ell}(X, \theta, G_0) = \frac{1}{\theta^2}\int \mu^2 dG(\mu) > 0 \tag{5.4}$$

unless $G = G_0$. Thus adaptation in the sense we have discussed is not possible. Note that the natural estimate $\bar{\theta}_n$ has asymptotic variance $2\theta^2/n$ in this case while $I^{-1}(\theta, G_0) = \theta^2/n$. Lindsay (1978, 1980) and Hammerstrom (1978) have independently studied situations such as this one (which are the rule rather than the exception) where adaptation is not possible. They have obtained what may be viewed as a minimax optimality property of $\bar{\theta}_n$ in Example 5 and analogous results in other problems of this type. We are investigating the natural extension of adaptation in this context.

5.2 *Better estimates.* The estimates we construct in Examples 2–4 have some serious failings: (i) the estimate of $\dot{\ell}$ is based on a small subsample rather than all the data; (ii) the estimates do not have natural invariance properties possessed by reasonable estimates in these problems, primarily because of the discretization of $\bar{\theta}_n$; and (iii) the behavior of the estimates when $I(\theta, G)$ is singular is not analyzed.

We believe that analogues of Stone's procedures in the location problem (which meet all these criticisms) can be constructed using the special structures of our examples. We have not pursued this since our interest lies primarily in illustrating the applicability of the general Condition H.

5.3 *Extensions to other asymptotic structures.* The theory we have developed extends naturally to cases where the observations are independent but not identically distributed, e.g., the usual linear model context. It can be applied, we believe, to the linear model and, as Stein's calculations and Wolfowitz (1974) indicate, to multiple regression models where both the location and the scale of the dependent variable are functions (possibly nonlinear) of the independent variables. Other extensions to non-independent situations, such as that treated in part in Beran (1976), should also be possible.

5.4 *Efficient estimation of functionals.* Levitt (followed by Ibragimov and Khazminski and others), in a series of papers starting with Levitt (1974), has studied how best to estimate functions $\theta(F)$ in nonparametric models, basing this work in part on Stein (1956). In some sense our problem can be viewed as the estimation of the solution $\theta(F)$ of $\int \dot{\ell}\,(x, \theta, G)\; dF_{(\theta,G)}(x) = 0$ which is meaningful (though possibly nonexistent) for $F \in \mathscr{F}$. Beyond this formal connection there seems to be no real link between our studies.

5.5 *Uniformity of adaptation.* Beran (1978) notes in the location problem (Example 1) that adaptive estimates converge to their limiting distributions uniformly on (shrinking n-dependent) "contiguous" neighborhoods of each G. This property can, we believe, be suitably re-expressed to apply generally. However, the weakness of this property is pointed out by Klaassen (1980) who shows (in Example 1, his Theorems 3.2.1 and 3.3.2) that for reasonable fixed neighborhoods the convergence is far from uniform. Thus from a practical point of view adaptive estimates may not work nearly as well for moderate samples as we might expect.

5.6 *Practical questions.* The difficulty of nonparametric estimation of score functions suggests that a more practical goal is partial adaptation, the construction of estimates which are (i) always $\sqrt{n}$-consistent, and (ii) efficient over a large parametric subfamily of $\mathscr{F}$. Our results indicate that when the orthogonality Condition S* and $\sqrt{n}$-consistency Condition GR(iv) hold, this goal should be achievable by using a one-step Newton approximation to the maximum likelihood estimate for the parametric subfamily by starting with an estimate which is $\sqrt{n}$-consistent for all of $\mathscr{F}$. Partial adaptation in Example 2 is discussed in Hogg (1980). This highlights an important practical and theoretical question in problems of this type, how to construct $\sqrt{n}$-consistent estimates. When there are no nuisance parameters present and adaptation is possible, maximum likelihood estimates for fixed shapes are natural candidates. In general, this question deserves further study. The constructions of Birgé (1980) may prove useful.

6. Theoretical Details.

6.1 *Proof of Lemma* 4.1. We use Stone's (1975) approach. Let ϕ_σ be the $\mathscr{N}(0, \sigma^2)$ density, g be any density, and define the convolution of g and ϕ_σ

$$g_\sigma = g * \phi_\sigma \tag{6.1}$$

and the convolution of the empirical d.f. and ϕ_σ

$$\hat{g}_\sigma(y) = m^{-1} \textstyle\sum_{i=1}^m \phi_\sigma(y - \varepsilon_i). \tag{6.2}$$

We suppress dependence on $\varepsilon_1, \cdots, \varepsilon_m$ in what follows.

For given $\sigma_m, c_m, d_m, e_m > 0$ define

$$q_m(y) = \begin{cases} \dfrac{\hat{g}'_{\sigma_m}}{\hat{g}}(y) & \text{if } \hat{g}_{\sigma_m}(y) \ge d_m, \quad |y| \le e_m \quad \text{and} \quad |\hat{g}'_{\sigma_m}(y)| \le c_m \hat{g}_{\sigma_m}(y), \\ 0 & \text{otherwise.} \end{cases} \tag{6.3}$$

We claim that if $c_m \to \infty$, $e_m \to \infty$, $\sigma_m \to 0$ and $d_m \to 0$ in such a way that

$$\sigma_m c_m \to 0, \tag{6.4}$$

$$e_m \sigma_m^{-3} = o(m), \tag{6.5}$$

then q_m satisfies the conclusions of Lemma 4.1. The argument proceeds by

LEMMA 6.1. *If the conditions of Lemma* 4.1 *hold and* q_m *satisfies* (6.3)–(6.5), *then*

$$\int_{[g>0]} \left\{ q_m(y) - \frac{g'_{\sigma_m}}{g_{\sigma_m}}(y) \right\}^2 g_{\sigma_m}(y)\, dy \to_P 0. \tag{6.6}$$

PROOF. We use the elementary estimates noted in Stone. For κ_i universal constants and all y,

$$\operatorname{Var} \hat{g}_\sigma^{(i)}(y) \le \kappa_i \sigma^{-(2i+1)} m^{-1} g_\sigma(y), \qquad i = 0, 1, \cdots. \tag{6.7}$$

Denote the conditions in (6.3) by A, B, C and the left-hand side of (6.6) by $I_1 + I_2$, where

$$I_1 = \int_{ABC} \left\{ \frac{\hat{g}'_{\sigma_m}}{\hat{g}_{\sigma_m}}(y) - \frac{g'_{\sigma_m}}{g_{\sigma_m}}(y) \right\}^2 g_{\sigma_m}(y)\, dy \tag{6.8}$$

$$I_2 = \int_{[ABC]^c} \frac{[g'_{\sigma_m}]^2}{g_{\sigma_m}}(y)\, dy. \tag{6.9}$$

Bound $E(I_1)$ by

$$2\left[\int_{ABC} g_{\sigma_m}^{-1}(y) E\{\hat{g}'_{\sigma_m}(y) - g'_{\sigma_m}(y)\}^2\, dy + \int_{ABC} c_m^2 g_{\sigma_m}^{-1}(y) E\{\hat{g}_{\sigma_m}(y) - g_{\sigma_m}(y)\}^2\, dy \right] = o(1) \tag{6.10}$$

by (6.7), (6.4) and (6.5). Bound

$$E(I_2) \le \int \frac{[g'_{\sigma_m}]^2}{g_{\sigma_m}}(y)[P\{|\hat{g}'_{\sigma_m}(y)| > c_m \hat{g}_{\sigma_m}(y)\} + P\{\hat{g}_{\sigma_m}(y) < d_m, g(y) > 0\} + I(|y| > e_m)]\, dy. \tag{6.11}$$

We claim that

$$\hat{g}_{\sigma_m}(y) \to g(y) \quad \text{in probability for all } y \quad \text{if} \quad m\sigma_m \to \infty, \tag{6.12}$$

$$\hat{g}'_{\sigma_m}(y) \to g'(y) \quad \text{in probability a.e. } y \quad \text{if} \quad m\sigma_m^3 \to \infty, \tag{6.13}$$

$$\int \frac{{g'_{\sigma_m}}^2}{g_{\sigma_m}}(y)\, dy \le \int \frac{g'^2}{g}(y)\, dy \quad \text{for all } m. \tag{6.14}$$

Evidently (6.12) and (6.13) imply that if $c_m \to \infty$ and $d_m \to 0$, then the two probabilities in (6.11) tend to 0 a.e. y, while (6.12)–(6.14) imply uniform integrability of ${g'_{\sigma_m}}^2/g_{\sigma_m}(y)$ and hence that

$$EI_2 \to 0. \tag{6.15}$$

Together (6.10) and (6.15) will establish Lemma 6.1. It remains to prove (6.12)-(6.14). Now by (6.7), for all y,

$$\hat{g}_{\sigma_m}(y) - g_{\sigma_m}(y) \to 0 \quad \text{in probability if} \quad m\sigma_m \to \infty, \tag{6.16}$$

$$\hat{g}'_{\sigma_m}(y) - g'_{\sigma_m}(y) \to 0 \quad \text{in probability if} \quad m\sigma_m^3 \to \infty. \tag{6.17}$$

Continuity of g and (6.16) imply (6.12). To prove (6.13) write (using the absolute continuity of g),

$$\begin{aligned}\int_{-\infty}^{\infty} | g'_{\sigma m}(y) - g'(y)| \, dy &= \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} (g'(y - \sigma_m x) - g'(y))\phi(x) \, dx \right| dy \\ &\le \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} | g'(y - \sigma_m x) - g'(y)| \, dy\phi(x) \, dx.\end{aligned} \tag{6.18}$$

Note that $I(G) < \infty$ implies $\int_{-\infty}^{\infty} |g'(y)| \, dy < \infty$. Thus we can apply the L_1 continuity theorem and the dominated convergence theorem to conclude that the right-hand side of (6.18) tends to 0 as $\sigma_m \to 0$ and (6.13) follows from (6.17) and (6.18). Finally, (6.14) is a well known inequality (see Hájek and Šidák, 1967, page 17). The lemma is proved. □

Next we need

LEMMA 6.2. *If* $\sigma \to 0$,

$$\int_{[g>0]} \left\{ \frac{g'_\sigma}{\sqrt{g_\sigma}}(y) - \frac{g'}{\sqrt{g}}(y) \right\}^2 dy \to 0. \tag{6.19}$$

PROOF. Apply (6.12)-(6.14).

LEMMA 6.3. *If* $\sigma_m c_m \to 0$,

$$\int_{[g>0]} q_m^2(y)(\sqrt{g_{\sigma_m}(y)} - \sqrt{g(y)}\,)^2 \, dy \to_P 0. \tag{6.20}$$

PROOF. Write, using Cauchy's form of Taylor's theorem,

$$\begin{aligned}\sqrt{g_\sigma(y)} - \sqrt{g(y)} &= \sigma \int_0^1 \left\{ \frac{\partial}{\partial\sigma} g_{\sigma\lambda}(y)/2g_{\sigma\lambda}^{1/2}(y) \right\} d\lambda \\ &= -\frac{\sigma}{2} \int_0^1 g_{\sigma\lambda}^{-1/2}(y) \int_{-\infty}^{\infty} zg'(y - \lambda\sigma z)\phi(z) \, dz \, d\lambda.\end{aligned} \tag{6.21}$$

Thus we can bound the square in the integrand of (6.20) by

$$\begin{aligned}&\frac{\sigma_m^2}{4} \int_0^1 g_{\lambda\sigma_m}^{-1}(y) \left\{ \int_{-\infty}^{\infty} zg'(y - \lambda\sigma_m z)\phi(z) \, dz \right\}^2 d\lambda \\ &\le \frac{\sigma_m^2}{4} \int_0^1 \int_{-\infty}^{\infty} \frac{\{zg'(y - \lambda\sigma_m z)\}^2}{g(y - \lambda\sigma_m z)} \phi(z) \, dz \, d\lambda\end{aligned} \tag{6.22}$$

by convexity of $(u, v) \to u^2/v$. Substitute (6.22) in (6.20) and use $| q_m | \le c_m$ to bound (6.20) by

$$\frac{c_m^2 \sigma_m^2}{4} \int_0^1 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{g'^2}{g}(v)\, z^2 \phi(z)\, dz\, dv\, d\lambda.$$

Since the integrals stay bounded independent of m, the result follows. □

Lemma 4.1 follows from Lemmas 6.1 and 6.3 since

$$\begin{aligned}
&\int \left\{ q_m(y) - \frac{g'}{g}(y) \right\}^2 g(y)\, dy \\
&\le 3 \Bigg[\int_{[g>0]} \left\{ q_m(y) - q_m \left(\frac{g_{\sigma_m}}{g} \right)^{1/2} (y) \right\}^2 g(y)\, dy \\
&\quad + \int_{[g>0]} \left\{ q_m \left(\frac{g_{\sigma_m}}{g} \right)^{1/2} (y) - \left(\frac{g'_{\sigma_m}}{g_{\sigma_m}} \right) \left(\frac{g_{\sigma_m}}{g} \right)^{1/2} (y) \right\}^2 g(y)\, dy \\
&\quad + \int_{[g>0]} \left\{ \left(\frac{g'_{\sigma_m}}{g_{\sigma_m}} \right) \left(\frac{g_{\sigma_m}}{g} \right)^{1/2} (y) - \frac{g'}{g}(y) \right\}^2 g(y)\, dy \Bigg],
\end{aligned} \tag{6.23}$$

and the first term tends to 0 by Lemma 6.3, the second by Lemma 6.1, and the last by Lemma 6.2. □

6.2 *Consistency Proofs.*

(i) *Consistency of $\hat{I}_n$ in* (3.14). As usual, we can take $\bar{\theta}_n$ to be deterministic, and in view of (3.15) we need only check that

$$\Delta_n = \bar{n}^{-1} \textstyle\sum_{i=m+1}^{n} \{ \hat{\dot{\ell}}^T \hat{\dot{\ell}}(X_i, \theta_n; X_1, \cdots, X_m) - \dot{\ell}^T \dot{\ell}(X_i, \theta_n, G) \} \to_{P_{\theta_n}} 0 \tag{6.24}$$

whenever $|\theta_n - \theta| = O(n^{-1/2})$. But by (3.11),

$$\begin{aligned}
&E_{\theta_n}\{ |\Delta_n| \,|X_1, \cdots, X_m \} \\
&\le E\{ |\hat{\dot{\ell}}^T \hat{\dot{\ell}}(X_{m+1}, \theta_n; X_1; X_1, \cdots, X_m) - \dot{\ell}^T \dot{\ell}(X_{m+1}, \theta_n, G)| \,X_1, \cdots, X_m \} \\
&= o_{P_{\theta_n}}(1)
\end{aligned} \tag{6.25}$$

and the result follows.

(ii) *Consistency in Theorem* 4.2. Again we can treat $\bar{\theta}_n$ as deterministic. Define measures $\{Q_n\}$ on $(R^{p+1})^n$ with densities

$$\textstyle\prod_{i=1}^{m} r(c_i) g(y_i - c_i \theta)^T \prod_{i=m+1}^{n} r(c_i) g(y_i - c_i(\theta - \bar{\theta}_n)^T).$$

We can argue as in the proof of (4.12) that the measures $\{Q_n\}$ are contiguous to the product measures specifying the distribution of the observations when θ is true. It follows that (4.28) is equivalent to

$$\bar{n}^{-1} \textstyle\sum_{i=m+1}^{n} q_m(\varepsilon_i; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m) = o_P(1). \tag{6.26}$$

By the usual calculation, conditioning on the first m observations,

$$\begin{aligned}
&E\left(\left[\bar{n}^{-1} \textstyle\sum_{i=m+1}^{n} \left\{ q_m(\varepsilon_i; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m) - \frac{g'}{g}(\varepsilon_i) \right\} \right]^2 \Bigg| \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m \right) \\
&= \int \left\{ q_m(y; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m) - \frac{g'}{g}(y) \right\}^2 g(y)\, dy = o_P(1)
\end{aligned}$$

by (4.13) and we can substitute g'/g for q_m in (6.26). With this final substitution, (4.28) follows from the WLLN. □

6.3 *Identities of Example* 4.

Verification of (4.31) *and* (4.32). Write $\dot{\ell} = (\dot{\ell}_1, \dot{\ell}_2)$ where

$$\dot{\ell}_1 = \left(\frac{\partial \ell}{\partial \mu_1}, \cdots, \frac{\partial \ell}{\partial \mu_k}\right), \qquad \dot{\ell}_2 = \left\{\frac{\partial \ell}{\partial v_{ij}} : i \geq j\right\}.$$

Evidently

$$\begin{aligned} \dot{\ell}_1(x, \theta, G_0) &= -\,|(x-\mu)V^{1/2}|^{-1}\frac{\gamma_0'}{\gamma_0}(|(x-\mu)V^{1/2}|)(x-\mu)V \\ &= \dot{\ell}_1((x-\mu)V^{1/2}, 0, [J], G_0)V^{1/2}, \end{aligned} \tag{6.27}$$

$$\dot{\ell}_2(x, \theta, G_0) = \left\{\left(\frac{(x_i-\mu_i)(x_j-\mu_j)}{|(x-\mu)V^{1/2}|}\frac{\gamma_0'}{\gamma_0}(|(x-\mu)V^{1/2}|) - v^{ij}\right)\left(1-\frac{\delta_{ij}}{2}\right)\right\}, \tag{6.28}$$

where $V^{-1} = \| v^{ij} \|$ and $x = (x_1, \cdots, x_k)$.

Define a linear operator L_B on $R^{k(k+1)/2}$, corresponding to a $k \times k$ matrix $B = \| b_{ij} \|$, by the $\dfrac{k(k+1)}{2} \times \dfrac{k(k+1)}{2}$ matrix

$$L_B = \left\| (b_{ir}b_{sj} + b_{jr}b_{is})\left(1 - \frac{\delta_{ij}}{2}\right) \right\|, \qquad r \geq s,\ i \geq j,$$

where (r, s) indexes rows and (i, j) columns. It is easy to verify that

$$\dot{\ell}_2(x, \theta, G_0) = \dot{\ell}_2((x-\mu)V^{1/2}, 0, [J], G_0)L_V^{-1/2}. \tag{6.29}$$

By (6.27) and (6.29) we have

$$I(\theta, G_0) = \begin{pmatrix} V^{1/2} & 0 \\ 0 & L_{V^{-1/2}} \end{pmatrix}^T I(0, [J], G_0)\begin{pmatrix} V^{1/2} & 0 \\ 0 & L_{V^{-1/2}} \end{pmatrix} \tag{6.30}$$

and, finally,

$$\dot{\ell}(x, \theta, G_0)I^{-1}(\theta, G_0) = \dot{\ell}((x-\mu)V^{1/2}, 0, [J], G_0)I^{-1}(0, [J], G_0) \times \begin{pmatrix} V^{1/2} & 0 \\ 0 & L^T_{V^{-1/2}} \end{pmatrix}^{-1}$$

Since $V^{1/2}$ is symmetric, (4.31) follows. To get (4.32) it is enough to verify that

$$L_B^{-1} = L_{B^{-1}} \quad \text{for any} \quad B, \tag{6.31}$$

and that if x is a triangular array

$$xL_B^T = [BQ(x)B^T], \tag{6.32}$$

where $Q(x)$ is the symmetric matrix whose ij-th entry is x_{ij} if $i \geq j$, or x_{ji} if $i < j$. The verifications of (6.31) and (6.32) are straightforward exercises in matrix multiplication.

Verification of (4.33) *and* (4.34). In this case $V^{1/2} = J$. For convenience suppress $(0, [J], G_0)$ in the arguments of functions for this discussion. We have

$$\dot{\ell}_1(x) = -\frac{x}{|x|}\frac{\gamma_0'}{\gamma_0}(|x|), \tag{6.33}$$

$$E\dot{\ell}_1^T\dot{\ell}_1(X) = E\left(\frac{\gamma_0'}{\gamma_0}\right)^2(|X|)\frac{X^TX}{|X|^2} = \frac{1}{k}\left\{E\left(\frac{\gamma_0'}{\gamma_0}\right)^2(|X|)\right\}J \tag{6.34}$$

by symmetry. Next, note that

$$\dot{\ell}_2(x) = \left\{\left(\frac{x_ix_j}{|x|}\frac{\gamma_0'}{\gamma_0}(|x|) - \delta_{ij}\right)(1 - \delta_{ij}/2)\right\}_{i \geq j} \tag{6.35}$$

and by symmetry

$$E\dot{\ell}_2^T\dot{\ell}_1(X) = 0 \tag{6.36}$$

$$E\dot{\ell}_2^T\dot{\ell}_2(X) = \| a_{rs,ij} \|_{r\geq s, i\geq j}, \tag{6.37}$$

where $X = (X_1, \cdots, X_k)$

$$a_{rs,ij} = 0, \quad \text{unless } r = i,\ s = j,$$

$$a_{rs,rs} = E\left\{\frac{X_1^2X_2^2}{|X|^2}\left(\frac{\gamma_0'}{\gamma_0}\right)(|X|)\right\}, \quad r \neq s, \tag{6.38}$$

$$a_{rr,rr} = E\left\{\frac{X_1^2}{|X|}\frac{\gamma_0'}{\gamma_0}(|X|) + 1\right\}^2.$$

$$E\left\{\frac{X_1^2X_2^2}{|X|^2}\left(\frac{\gamma_0'}{\gamma_0}\right)^2(|X|)\right\} = E\left\{\frac{X_1^2X_2^2}{|X|^4}\right\}E\left\{|X|^2\left(\frac{\gamma_0'}{\gamma_0}\right)^2(|X|)\right\} \tag{6.39}$$

by spherical symmetry of G_0. The second term in (6.39) is just $I_2(G_0)$, while the first term is independent of G_0 and may be shown to equal $k^{-1}(k+2)^{-1}$ by taking G_0 to be the spherical normal distribution. Thus

$$a_{rs,rs} = k^{-1}(k+2)^{-1}I_2(G_0), \quad r \neq s. \tag{6.40}$$

A similar computation gives

$$a_{rr,rr} = \frac{1}{4}E\left\{\frac{X_1^4}{|X|^2}\left(\frac{\gamma_0'}{\gamma_0}\right)^2(|X|)\right\} - 1 = \frac{1}{4}3k^{-1}(k+2)^{-1}I_2(G_0) - 1. \tag{6.41}$$

We see from (6.37) that $I(0, [J], G_0)$ is a diagonal matrix with entries given by (6.40) and (6.41). Upon inverting it and substituting (6.40) and (6.41) in $\dot{\ell}(x, 0, [J], G_0)$, we obtain (4.33) and (4.34).

6.4 *Two Theorems on efficient estimates.*

THEOREM 6.1. *Under R suppose $\{\hat{\theta}_n\}$ are such that, for a given θ, $\mathscr{L}_{\theta_n}\{n^{1/2}(\hat{\theta}_n - \theta_n)\} \to \mathscr{N}(0, I^{-1}(\theta))$ whenever $n^{1/2}|\theta_n - \theta| \leq M$ for all n, $M < \infty$. Then,*

$$n^{1/2}(\hat{\theta}_n - \theta) = n^{-1/2}\textstyle\sum_{i=1}^n \dot{\ell}(X_i, \theta)I^{-1}(\theta) + o_{p_\theta}(1). \tag{6.42}$$

NOTE. This claim is in fact valid in great generality if the local asymptotic normality (LAN) condition of Hájek (1972) holds with $\Delta_n(\theta)$ replacing $n^{-1/2}\sum_{i=1}^n \dot{\ell}(X_i, \theta)$. Moreover it is clear that everything is local so that the condition and conclusion need only hold at a point θ on which $\hat{\theta}_n$ can depend.

PROOF. Since the sequence of joint laws $\mathscr{L}_n$ of $n^{1/2}(\hat{\theta}_n - \theta)$ and $n^{-1/2}\sum_{i=1}^n \dot{\ell}(X_i, \theta)I^{-1}(\theta)$ is tight under P_θ it is enough to show that if $\mathscr{L}_{m_n}$ is any subsequence weakly convergent to $\mathscr{L}^*$ (say) then $\mathscr{L}^*$ must concentrate on the diagonal. by a contiguity and analyticity argument, see Roussas (1972, pages 136-141), we can show that the joint characteristic function $\phi^*(u, v)$ of $\mathscr{L}^*$ satisfies the equation

$$\phi^*(u, v) = \phi^*(u, 0)\exp\{-uI^{-1}(\theta)v^T\}\exp\{-\tfrac{1}{2}vI^{-1}(\theta)v^T\}$$

(Substitute $\Gamma = I(\theta)$, $h = vI^{-1}(\theta)$ in (3.11) of Roussas.) But, by hypothesis,

$$\phi^*(u, 0) = \exp\{-\tfrac{1}{2}uI^{-1}(\theta)u^T\}$$

so that

$$\phi^*(u, v) = \exp\{-\tfrac{1}{2}(u+v)I^{-1}(\theta)(u+v)^T\},$$

and the theorem follows. □

THEOREM 6.2. *If* R(i), R(ii) *and* UR(iii) *hold and if* $\bar{\theta}_n$ *is* $\sqrt{n}$*-consistent and discretized as in* (2.3) *and*

$$\hat{\theta}_n = \bar{\theta}_n + n^{-1} \textstyle\sum_{j=1}^n \dot{\ell}(X_j, \bar{\theta}_n) I^-(\bar{\theta}_n),$$

then $\hat{\theta}_n$ *is efficient in the usual sense.*

PROOF. In view of the arguments leading to Theorem 4 of Le Cam (1968), it is enough to show that for θ regular and any sequence θ_n such that $n^{1/2} \mid \theta_n - \theta \mid \leq M$ for all n

$$n^{-1/2} \textstyle\sum_{i=1}^n \{\dot{\ell}(X_i, \theta_n) - \dot{\ell}(X_i, \theta)\} + n^{1/2}(\theta_n - \theta) I(\theta) = o_{P_\theta}(1). \tag{6.43}$$

We claim that (6.43) is implied by the fact that

$$\textstyle\sum_{i=1}^n \{\ell(X_i, \theta_n + hn^{-1/2}) - \ell(X_i, \theta_n)\} = hn^{-1/2} \sum_{i=1}^n \dot{\ell}(X_i, \theta_n) - \tfrac{1}{2} h I(\theta_n) h^T + o_{P_\theta}(1) \tag{6.44}$$

for all h. To see this, note that from the usual LAN condition

$$\begin{aligned}\textstyle\sum_{i=1}^n \{\ell(X_i, \theta_n + hn^{-1/2}) - \ell(X_i, \theta)\} &= n^{1/2}(\theta_n - \theta) + hn^{-1/2} \textstyle\sum_{i=1}^n \ell(X_i, \theta) \\ &\quad -\tfrac{1}{2}\{n^{1/2}(\theta_n - \theta) + h\} I(\theta) \{n^{1/2}(\theta_n - \theta) + h\}^T + o_{P_\theta}(1);\end{aligned} \tag{6.45}$$

$$\begin{aligned}\textstyle\sum_{i=1}^n \{\ell(X_i, \theta_n) - \ell(X_i, \theta)\} &= n^{1/2}(\theta_n - \theta) n^{-1/2} \textstyle\sum_{i=1}^n \dot{\ell}(X_i, \theta) \\ &\quad - \frac{n}{2}\{(\theta_n - \theta) I(\theta)(\theta_n - \theta)^T\} + o_{P_\theta}(1).\end{aligned} \tag{6.46}$$

Subtracting (6.46) from (6.45) and matching the coefficient of h in (6.44) yields (6.43).

Finally, (6.44) is just the usual statement of LAN with θ replaced by θ_n. It is argued in exactly the same way as the usual equivalence,—see pages 54–63 of Roussas (1972) for example,—but, of course, we use the uniformity in UR(iii). The theorem follows. □

Acknowledgement V. Fabian and J. Hannan corrected my mistaken impression that R(i) − R(iii) were sufficient to establish the efficiency of $\bar{\theta}_n + n^{-1/2} \sum \dot{\ell}(X_i, \bar{\theta}_n)$. I am grateful to them for prompting me to prove Theorems 6.1 and 6.2 as well as other valuable comments. I am also grateful to Chris A. J. Klaassen for a careful reading of the paper resulting in several substantial corrections and to J. Pfanzagl for the Ibragimov-Hasminskii reference.

REFERENCES

BERAN, R. (1974). Asymptotically efficient adaptive rank estimates in location models. *Ann. Statist.* **2** 63–74.

BERAN, R. (1975). Adaptive estimates for autoregressive processes. *Ann. Inst. Statist. Math.* **28** 77–89.

BERAN, R. (1978). An efficient and robust adaptive estimator of location. *Ann. Statist.* **6** 292–313.

BICKEL, P. J. (1981). Lectures on robustness and adaptation, to appear in 1979 *St. Flour Conference Lecture Notes in Mathematics* 876 Springer, Berlin.

BIRGÉ, L. (1980). Approximation dans les espaces métrique et théorie de l'estimation. Unpublished thesis, University of Paris.

DIONNE, L. (1981). Efficient nonparametric estimators of parameters in the general linear hypothesis. *Ann. Statist.* **9** 457–460.

FABIAN V. and HANNAN J. (1980). On estimation and adaptive estimation for locally asymptotically normal families. *Z. Wahrsch. verw. Gebiete.* To appear.

HÁJEK, J. (1962). Asymptotically most powerful rank order tests. *Ann. Math. Statist.* **33** 1124–1147.

HÁJEK, J. and SIDÁK, Z. (1967). *Theory of Rank Tests*, Academic, New York and Academia, Prague.

HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* **1** 175–194. University of California Press, Berkeley.

HAMMERSTROM, T. (1978). Ph.D. Thesis, University of California, Berkeley.

HASMINSKII, P. Z. and IBRAGIMOV I. A. (1978). On the nonparametric estimation of functionals. *Proc. Second Prague Symp. Asympt. Statistics and Probability*, J. Jureçkova Ed., Prague.

HOGG, R. (1980). On adaptive robust inference. Tech. Report, Univ. of Iowa.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings Fifth Berkeley Symp. Math. Statist. Prob.* **1** 221–233 University of California Press, Berkeley.

HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Math. Statist.* **1** 799–821.

HUBER, P. J. (1977). Robust covariances *Statistical Decision Theory and Related Topics*, II. S. S. Gupta and D. S. Moore, eds. 165–192 Academic, New York.

KLAASSEN, C. (1980). Statistical performance of location estimators. Thesis, University of Leiden, Netherlands.

LE CAM, L. (1969). *Théorie Asymptotique de la Décision Statistique*, Les Presses de l'Univérsite de Montreal.

LEVITT, B. (1974). On optimality of some statistical estimates, *Proc. Prague Symp. Asympt. Statist.* 215–238 J. Jureckova Ed. Prague.

LINDSAY, B. (1978). Information in the presence of nuisance parameters. Thesis, University of Washington, Seattle.

LINDSAY, B. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood. *Phil. Trans. Roy. Soc. London* **296** 639–665.

MARONNA, R. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67.

ROUSSAS, G. (1972). *Contiguity of Probability Measures: Applications in Statistics.* Cambridge University Press.

STEIN, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 187–196. University of California Press.

STONE, C. (1975). Adaptive maximum likelihood estimation of a location parameter. *Ann. Statist.* **3** 267–284.

TAKEUCHI, K. (1971). A uniformly asymptotically efficient estimator of a location parameter, *J. Amer. Statist. Assoc.* **66** 292–301.

VAN EEDEN, C. (1970). Efficiency-robust estimation of location. *Ann. Math. Statist.* **41** 172–181.

WEISS, L. and WOLFOWITZ, J. (1970). Asymptotically efficient nonparametric estimators of location and scale parameters. *Z. Wahrsch. verw. Gebiete* **16** 134–150.

WEISS, L. and WOLFOWITZ, J. (1971). Asymptotically efficient estimation of nonparametric regression coefficients, *Statistical Decision Theory and Related Topics*, S. Gupta and J. Yackel, 29–40 eds. Academic, New York.

WOLFOWITZ, J. (1953). The method of maximum likelihood and the Wald theory of decision functions. *Indag. Mathemat.* **56** 114–119.

WOLFOWITZ, J. (1974). Asymptotically efficient nonparametric estimators of location and scale parameters. II. *Z. Wahrsch. verw. Gebiete* **30** 117–128.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

ADVANCES IN APPLIED MATHEMATICS 7, 55–69 (1986)

Empirical Bayes Estimation in Functional and Structural Models, and Uniformly Adaptive Estimation of Location

P. J. BICKEL*

Department of Statistics, University of California, Berkeley, California 94720

AND

C. A. J. KLAASSEN*,†

Department of Mathematics, University of Leiden, Postbus 9512, 2300 RA Leiden, Netherlands

DEDICATED TO HERBERT ROBBINS ON THE OCCASION OF HIS 70TH BIRTHDAY

We discuss estimation of parameters in functional and structural models in relation to Robbins' empirical Bayes and compound decision theories. We construct an efficient estimate of ν in the normal functional model, X_i independent $\mathcal{N}(\nu, \theta_i)$ where $\varepsilon \le \theta_i^2 \le 1/\varepsilon$, $\varepsilon > 0, 1 \le i \le n$.

1. INTRODUCTION

In 1956, Robbins [15] (see also Good [4]) initiated the systematic study of nonparametric empirical Bayes procedures. Robbins [16] is a good entry to the large literature. The focus of his work and that of its many successors has been the model:

I: We observe random variables or vectors $X_1, \dots, X_n$ i.i.d. F where F ranges over all (or most) mixtures of a parametric family $\{F_\theta : \theta \in \Theta\}$ with

*Research partially supported by ONR Contract N00014-80-C-0163.

†Research carried out in part with the support of the Mathematical Sciences Research Institute (Berkeley).

0196-8858/86 $7.50

$\Theta \subset R^p$. That is,

$$F = \int F_\theta dG(\theta)$$

for some probability G on Θ, belonging to a set $\mathscr{G}$. Equivalently, we observe X_i, $1 \le i \le n$ where (θ_i, X_i) are i.i.d. with $\theta_i \sim G$ and given θ_i, $X_i \sim F_{\theta_i}$. Work in the area has focused on questions such as simultaneous estimation of $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^T$ with squared error loss, $L(\boldsymbol{\theta}, \mathbf{d}) = n^{-1}\sum_{i=1}^n(\theta_i - d_i)^2$, $\mathbf{d} = (d_1, \dots, d_n)^T$, and the possibility of constructing decision rules

$$\delta^*(\mathbf{X}) = \left(h^*(X_1; \mathbf{X}), \dots, h^*(X_n; \mathbf{X})\right)^T, \qquad \mathbf{X} = (X_1, \dots, X_n)^T, \tag{1.1}$$

which to first order approximate the Bayes rule,

$$\delta(\mathbf{X}, G) = \left(h(X_1, G), \dots, h(X_n, G)\right)^T$$

where

$$h(X, G) = E(\theta | X), \qquad (\theta, X) \sim (\theta_1, X_1).$$

Robbins came to the empirical Bayes formulation from his 1951 consideration of the compound decision problem [14].

II: Observe X_i independent with $X_i \sim F_{\theta_i}$, $\theta_i \in K$ compact $\subset \Theta$, for $1 \le i \le n$. A typical problem now is to simultaneously estimate $\theta_1, \dots, \theta_n$ as well as possible, asymptotically, i.e., to find $\delta_n^*(X_1, \dots, X_n) = (h_{1n}^*(\mathbf{X}), \dots, h_{nn}^*(\mathbf{X}))^T$ such that

$$\liminf_n n^{-1} \sum_{i=1}^n \left\{ E_{\theta_i}\left(h_{in}(\mathbf{X}) - \theta_i\right)^2 - E_{\theta_i}\left(h_{in}^*(\mathbf{X}) - \theta_i\right)^2 \right\} \le 0 \tag{1.2}$$

for any competing sequence $\delta_n(\mathbf{X}) = (h_{1n}(\mathbf{X}), \dots, h_{nn}(\mathbf{X}))^T$. The solution, heuristically, is to use δ^* given by (1.1) since the risks in (1.2) should be close to model I risks when $G = G_n$ is the empirical distribution of $\theta_1, \dots, \theta_n$.

A key element in the transition from I to II evidently lies in establishing that the approximation of $\delta(\mathbf{X}, G)$ by $\delta^*(\mathbf{X})$ is in a suitable sense, uniform in G.

An analogous set of questions was investigated by Neyman and Scott [12], Kiefer and Wolfowitz [8], and notably, recently Lindsay [10] and others. Their focus is on estimating a parameter ν common to the X_i in the presence of random (structural models) or fixed (functional models) nuisance parameters $\theta_1, \ldots, \theta_n$. The corresponding models are:

I′: (Structural) $X_1, \ldots, X_n$ i.i.d. F where

$$F = \int F_{(\nu, \theta)} dG(\theta) \tag{1.3}$$

$G \in \mathscr{G}$, $\nu \in H$ open $\subset R^m$.

II′: (Functional) X_i independent with $X_i \sim F_{(\nu, \theta_i)}$, $\theta_i \in K \subset \theta$, K compact, $1 \leq i \leq n$.

Again $\{F_{(\nu, \theta)} : \nu \in H, \theta \in \Theta\}$ is a postulated parametric model.

In various examples discussed by these authors it is clear that ν can be estimated at rate $n^{-1/2}$. For instance, if $F_{(\nu, \theta)}$ is the $\mathscr{N}(\nu, \theta^2)$ distribution, $\bar{X}$ is a $n^{1/2}$ consistent estimate of ν in model I′ if $\int \theta^2 dG(\theta) < \infty$ and in II′ if the empirical second moment of θ, $n^{-1}\Sigma_{i=1}^n \theta_i^2$ is bounded. What are optimal procedures in this context? For simplicity take $m = 1$.

Let $F_{(\nu, G)}$ denote the distribution (1.3), $P_{(\nu, G)}$ the associated probability measure, etc. Call a (sequence of) estimate(s) *regular* (I′) if

$$\mathscr{L}_{(\nu_n, G_n)}\left(n^{1/2}(T_n - \nu_n)\right) \to \mathscr{N}\left(0, \sigma_T^2(\nu_0, G_0)\right) \tag{1.4}$$

whenever $\nu_n \to \nu_0$ and $G_n \to G_0$ (weakly) for all $\nu_0 \in H, G_0 \in \mathscr{G}$. Call T_n^* *efficient* (I′) if $\{T_n^*\}$ is regular (I′) and

$$\sigma_T^{2*}(\nu_0, G_0) \leq \sigma_T^2(\nu_0, G_0) \tag{1.5}$$

for all regular $\{T_n\}$, (ν_0, G_0).

In model I′ let $\mathscr{G}$ be the set of all probability distributions on K. Call an estimate *regular* (II′) if

(i) $T_n(x_1, \ldots, x_n)$ is symmetric in $(x_1, \ldots, x_n)$

(ii) $\mathscr{L}_{(\nu_n, \theta_1, \ldots, \theta_n)}(n^{1/2}(T_n - \nu_n)) \to \mathscr{N}(0, \sigma_T^2(\nu_0, G_0))$

whenever $\nu_n \to \nu_0$ and G_n, the empirical distribution, $n^{-1}\Sigma_{i=1}^n I(\theta_i \leq \cdot)$, of $\{\theta_1, \ldots, \theta_n\}$, tends (weakly) to $G_0 \in \mathscr{G}$. An estimate T_n^* is *efficient* (II′) if it is regular (II′) and satisfies (1.5) for regular (II′) competitors T_n.

In problem I′ sufficiency of the order statistics permits us to restrict to symmetric estimates. In problem II′ invariance of the problem under permutations of the θ_i leads less forcefully to the same conclusion. The passage from efficiency (I′) to efficiency (II′) is as in Robbins' problems a question of uniformity.

Evidently,

PROPOSITION 1.1. *Suppose $\mathscr{G}$ for both models is the set of all distributions on K.*

(i) *if T_n is regular* (II′) *it is regular* (I′)

(ii) *If T_n^* is efficient* (I′) *and regular* (II′) *then T_n^* is efficient* (II′).

An extension of the theory of information (Cramér–Rao) bounds to models with infinite dimensional nuisance parameters such as I′ has been developed by Koshevnik and Levit [9], Pfanzagl [13], and Begun *et al.* [1] on the basis of a fundamental paper of Stein [17]. Under regularity conditions, efficient (I′) estimates are regular (I′) estimates achieving these information bounds. Methods for constructing such estimates in a general context are discussed in [13, 2, 3] among others. We do not study the general situation further but show in an important special case how to construct estimates which are not only efficient (I′) but also regular (II′) and hence efficient (II′).

The example we consider and extend somewhat is the normal location problem with variances possibly changing from observation to observation.

$$F_{(\nu,\theta)} = \mathscr{N}(\nu, \theta^2) \tag{1.6}$$

with $\Theta = R^+$. Take $K = [\varepsilon, 1/\varepsilon]$ for fixed $\varepsilon > 0$ and $\mathscr{G}$, all distributions on K. Then $F_{(\nu,G)}$ is still a symmetric location family in ν. If G is known, efficient estimates are asymptotically $\mathscr{N}(\nu, I^{-1}(H)/n)$ where $H = \int F_{(0,\theta)} dG(\theta)$,

$$I(H) = \int \frac{[h']^2}{h}(t)\, dt$$

$$h(t) = \int_0^\infty \theta^{-1}\varphi(t\theta^{-1})\, dG(\theta).$$

The general information bound theory indicates that it should be possible to adapt perfectly in this case, i.e., do as well not knowing G as knowing it. In fact, Stone [18] constructs an estimate $\hat{\nu}_n$ which is location and scale equivariant and such that,

$$\mathscr{L}_F\big(n^{1/2}(\hat{\nu}_n - \nu)\big) \to \mathscr{N}\big(0, I^{-1}(H)\big) \tag{1.7}$$

whenever $X_1, \ldots, X_n$ are i.i.d. F and

$$F(\cdot) = F(\cdot + \nu) \text{ is symmetric about } 0. \tag{1.8}$$

Here, we define generally for H on $[-\infty, \infty] = \bar{R}$, $H(R) > 0$

$$
\begin{aligned}
I(H) &= \int_R \frac{[h']^2}{h}(t)\,dt \qquad \text{if } H \text{ has an absolutely continuous density } h \text{ on } R, \\
&= \infty \qquad \text{otherwise.}
\end{aligned}
\tag{1.9}
$$

For convenience, in the sequel, distribution functions are defined by capital letters and their densities, by convention, are the corresponding lower case letters. In Section 2 of this paper we construct a modified and simplified translation but not scale equivariant version of Stone's estimate, ν_n^*, which satisfies (1.7) and is also regular (II′) for the model (1.6). In fact, we show for the symmetric location model,

THEOREM 1.1. *$\mathscr{L}_{H_n}(n^{1/2}\nu_n^*)$ converges to $\mathscr{N}(0, I^{-1}(H_0))$ whenever*

(a) $H_n \xrightarrow{w} H_0$, $H_0(R) = 1$

(b) $I(H_n) \to I(H_0) < \infty$.

Then, in Theorem 2.1, we show that uniformity of convergence persists in a generalization of model II′.

Theorem 1.1 is the best that one can hope for in adaptive estimation of location since $\mathscr{L}_{H_m}(n^{1/2}\hat{\nu}_n) \to \mathscr{N}(0, I^{-1}(H_m))$ as $n \to \infty$, uniformly in m, $H_m \xrightarrow{w} H_0$ as $m \to \infty$, and $\sup_m I(H_m) < \infty$ imply that $I(H_m) \to I(H_0)$.

This estimate is also asymptotically minimax in Huber's [6] sense and can be used for the construction of an adaptive confidence interval, $\nu_n^* \pm z(nI_n^*)^{-1/2}$ where, $\inf_{\mathscr{H}} P_H[\nu_n^* - z(nI_n^*)^{-1/2} \le \nu \le \nu_n^* + z(nI_n^*)^{-1/2}] \to 2\Phi(z) - 1$ for any family $\mathscr{H}$ of distributions symmetric about 0 which does not have point mass at $\pm\infty$ as a weak limit point. The details of these results and other robustness properties of $\{\nu_n^*\}$ will appear in Bickel *et al.* [3].

2. The Results

Suppose the common distribution of $X_1, \ldots, X_n$ i.i.d. is H as in (1.8), with $H \in \mathscr{H}$. Suppose $\mathscr{H}$ does not have point mass at $\pm\infty$ as a weak limit point. Then there exist uniformly $n^{1/2}$ consistent translation equivariant estimates $\tilde{\nu}_n$ of ν, such that,

$$
\mathscr{L}_H\left(n^{1/2}(\tilde{\nu}_n - \nu)\right) \to \mathscr{N}\left(0, \sigma^2(H)\right) \tag{2.1}
$$

uniformly on $\mathscr{H}$ and $\sup_{\mathscr{H}} \sigma^2(H) < \infty$. For instance let

$$k(t) = \frac{e^{-t}}{(1 + e^{-t})^2} \tag{2.2}$$

be the logistic density. If $\tilde{\nu}_n$ is the unique solution of

$$\sum_{i=1}^{n} \frac{k'}{k}(X_i - \nu) = 0$$

then it is easy to see that $\tilde{\nu}_n$ satisfies (2.1).

To define ν_n^* we proceed as in Stone [18], but use the logistic rather than the normal kernel for smoothing. Let

$$k_\sigma(x) = \frac{1}{\sigma} k\left(\frac{x}{\sigma}\right).$$

If $\hat{F}_n$ is the empirical d.f. of $X_1, \ldots, X_n$, define

$$\hat{h}_\sigma(x) = \int k_\sigma(x - z)\, d\hat{F}_n(z) = \frac{1}{n} \sum_{i=1}^{n} k_\sigma(x - X_i).$$

Next let

$$\hat{q}_\sigma(x) = \frac{\hat{h}'_\sigma}{\hat{h}_\sigma}(x),$$

$$\bar{q}_\sigma(x, \nu) = \frac{1}{2}\left[\hat{q}_\sigma(x + \nu) - \hat{q}_\sigma(-x + \nu)\right].$$

Let ψ be symmetric and continuous at 0 with support $[-1, 1]$, $0 \le \psi \le 1$ and $\psi(0) = 1$. Let

$$\psi_n(x) = \psi(c_n x)$$

and $\sigma_n \downarrow 0$, $c_n \downarrow 0$ at a rate to be determined later. Write $\hat{q}_n, \bar{q}_n, \hat{h}_n$ for $\hat{q}_{\sigma_n}, \bar{q}_{\sigma_n}, \hat{h}_{\sigma_n}$. Then we define

$$\nu_n^*(\nu) = \nu - \hat{I}_n^{-1}(\nu) \int \bar{q}_n(x, \nu) \psi_n(x) \hat{h}_n(x + \nu)\, dx$$

where

$$\hat{I}_n(\nu) = \int \bar{q}_n^2(x, \nu) \psi_n(x) \hat{h}_n(x + \nu)\, dx.$$

Finally our estimate is

$$\nu_n^* = \nu_n^*(\tilde{\nu}_n).$$

Since we have selected $\tilde{\nu}_n$ to be translation equivariant, the second term of ν_n^* is translation invariant and ν_n^* itself is translation equivariant, and therefore we may and do assume that the true value of $\nu = 0$, i.e., that H_n is the common distribution of the X_i. We then define the density and score function of the convolution $\tilde{H}_n$ of H_n with the logistic distribution with mean 0 and variance σ_n^2

$$\tilde{h}_n(x) = \int k_{\sigma_n}(x - z) h_n(z)\, dz$$

$$\tilde{q}_n(x) = \frac{\tilde{h}'_n}{\tilde{h}_n}(x).$$

Then,

$$\tilde{h}_n(x) = E\hat{h}_n(x) \tag{2.3}$$

and $\hat{I}_n(\nu_n^*)$ estimates the quantity

$$I_n(\tilde{H}_n) = \int \tilde{q}_n^2(x)\psi_n(x)\tilde{h}_n(x)\, dx.$$

We prove Theorem 1.1 by a series of lemmas. The proof is somewhat simpler than our original thanks to an idea of J. Ritov. Uniformly for $H_n \in \mathscr{H}$,

LEMMA 2.1. *Write $\bar{q}_n(x)$ for $\bar{q}_n(x, 0)$ etc. Then,*

$$\begin{aligned}
&\int [\bar{q}_n(x,\nu)\psi_n(x)\hat{h}_n(x+\nu) - \bar{q}_n(x)\psi_n(x)\hat{h}_n(x)]\, dx \\
&\quad - \nu\int \bar{q}_n(x)\psi_n(x)\hat{h}'_n(x)\, dx \\
&= 0_p\left(\nu\int(\hat{q}'_n(x) - \hat{q}'_n(-x))\psi_n(x)\hat{h}_n(x)\, dx\right) + 0_p(\sigma_n^{-3}\nu^2)
\end{aligned} \tag{2.4}$$

$$\hat{I}_n(\nu) = \hat{I}_n + 0_p(\sigma_n^{-3}\nu). \tag{2.5}$$

LEMMA 2.2.

$$\int(\hat{q}'_n(x) - \hat{q}'_n(-x))\psi_n(x)\hat{h}_n(x)\, dx = 0_p(\sigma_n^{-3}c_n^{-1}n^{-1}) \tag{2.6}$$

$$\int \bar{q}_n(x)\psi_n(x)\hat{h}'_n(x)\, dx = \hat{I}_n + 0_p(\sigma_n^{-3}c_n^{-1}n^{-1}). \tag{2.7}$$

LEMMA 2.3.

$$\int \bar{q}_n(x)\psi_n(x)\hat{h}_n(x)\,dx = \int \tilde{q}_n(x)\psi_n(x)\hat{h}_n(x)\,dx + 0_p\left(n^{-1}c_n^{-1}\sigma_n^{-2}\right) \tag{2.8}$$

$$\hat{I}_n = I_n(\tilde{H}_n) + 0_p\left(\sigma_n^{-5/2}c_n^{-1/2}n^{-1/2}\right) \tag{2.9}$$

LEMMA 2.4. *If* $c_n = \sigma_n$, $n\sigma_n^6 \to \infty$, *and* $\sup_n I(H_n) < \infty$, *then*

$$n^{1/2}\nu_n^* = -n^{-1/2}I_n^{-1}(\tilde{H}_n)\sum_{i=1}^{n}\int \tilde{q}_n(x)\psi_n(x)k_{\sigma_n}(x - X_i)\,dx + o_p(1). \tag{2.10}$$

LEMMA 2.5. *If* $H_n \xrightarrow{w} H_0$ *and* $\sup_n I(H_n) < \infty$,

$$\liminf_n I_n(H_n) \geq I(H_0) \tag{2.11}$$

and

$$\liminf_n I(H_n) \geq \liminf_n I_n(\tilde{H}_n) \geq I(H_0). \tag{2.12}$$

If also $I(H_n) \to I(H_0) < \infty$, *then*

$$\int \left(h_n^{-1/2}h_n' - h_0^{-1/2}h_0'\right)^2(x)\,dx \to 0. \tag{2.13}$$

LEMMA 2.6. *If* $H_n \xrightarrow{w} H_0$, $H_0(R) = 1$, *and* $I(H_n) \to I(H_0) < \infty$, *the family of product measures* Q_n, θ *with density* $\{\pi_{i=1}^n h_n(x_i - \theta/n^{1/2})\}$ *satisfies Le Cam's L.A.N. condition and*

$$\log\frac{dQ_{n,\theta}}{dQ_{n,0}} = \theta n^{-1/2}\sum_{i=1}^{n}\frac{h_n'}{h_n}(X_i) - \frac{1}{2}\theta^2 I(H_n) + o_p(1) \tag{2.14}$$

and

$$\mathscr{L}_{H_n}\left(n^{-1/2}\sum_{i=1}^{n}\frac{h_n'}{h_n}(X_i)\right) \to \mathscr{N}(0, I(H_0)).$$

Proof of Lemma 2.1. Taylor expand, about $\nu = 0$, to find that (2.4) equals

$$\frac{\nu}{2}\int \left(\hat{q}_n'(x) - \hat{q}_n'(-x)\right)\psi_n(x)\hat{h}_n(x)\,dx$$
$$+\nu^2\int\int_0^1 (1-\lambda)\left[\frac{\partial^2}{\partial\mu^2}\left(\bar{q}_n(x,\mu)\hat{h}_n(x+\mu)\right)\Big|_{\mu=\lambda\nu}\right]\psi_n(x)\,d\lambda\,dx.$$

Note that if $\|\cdot\|$ is the sup norm,

$$\left\|\frac{\hat{h}_n^{(r)}}{\hat{h}_n}\right\| = 0_p(\sigma_n^{-r}), \qquad \left\|\frac{\tilde{h}_n^{(r)}}{\tilde{h}_n}\right\| = 0_p(\sigma_n^{-r}),$$

since there exist finite constants C_r with

$$\left|\int k^{(r)}(x)\,dG(x)\right| \le C_r \int k(x)\,dG(x) \qquad \text{for all } r, G. \tag{2.15}$$

Hence,

$$\left\|\frac{\partial^r \bar{q}_n}{\partial \nu^r}(\cdot, \nu)\right\| = 0_p\left(\sum_{s=1}^{r+1}\left\|\frac{\hat{h}_n^{(s)}}{\hat{h}_n}\right\|^{r+1/s}\right)$$
$$= 0_p\left(\sigma_n^{-(r+1)}\right)$$

and (2.4) follows. A similar argument yields (2.5).

Proof of Lemma 2.2. Write, using symmetry,

$$\int (\hat{q}_n'(x) - \hat{q}_n'(-x))\psi_n \hat{h}_n(x)\,dx$$
$$= \int (\hat{q}_n'(x) - \hat{q}_n'(-x))\psi_n(\hat{h}_n - \tilde{h}_n)(x)\,dx$$
$$= 0_p\left(\left[\int (\hat{q}_n' - \tilde{q}_n')^2 \psi_n \tilde{h}_n(x)\,dx\right]^{1/2} \times \left(\int \frac{(\hat{h}_n - \hat{h}_n)^2}{\tilde{h}_n}\psi_n(x)\,dx\right)^{1/2}\right). \tag{2.16}$$

By (2.15),

$$E(\hat{h}_n^{(r)} - \tilde{h}_n^{(r)})^2(x) \le \tfrac{1}{4}C_r^2 n^{-1}\sigma_n^{-(2r+1)}\tilde{h}_n(x) \tag{2.17}$$

and consequently

$$\int (\hat{h}_n^{(r)} - \tilde{h}_n^{(r)})^2 \tilde{h}_n^{-1}\psi_n(x)\,dx = 0_p\left(n^{-1}\sigma_n^{-(2r+1)}c_n^{-1}\right). \tag{2.18}$$

Next write

$$(\hat{q}_n' - \tilde{q}_n')^2 \le 2\left\{\left(\frac{\hat{h}_n''}{\hat{h}_n} - \frac{\tilde{h}_n''}{\tilde{h}_n}\right)^2 + \left(\hat{q}_n^2 - \tilde{q}_n^2\right)^2\right\} \tag{2.19}$$

$$\frac{\hat{h}_n''}{\hat{h}_n} - \frac{\tilde{h}_n''}{\tilde{h}_n} = \frac{\hat{h}_n''}{\hat{h}_n}\left(\frac{\tilde{h}_n - \hat{h}_n}{\tilde{h}_n}\right) + \tilde{h}_n^{-1}(\hat{h}_n'' - \tilde{h}_n'') \tag{2.20}$$

$$\begin{aligned}|\hat{q}_n^2 - \tilde{q}_n^2| &= |\hat{q}_n + \tilde{q}_n||\hat{q}_n - \tilde{q}_n| \\ &\le 2\sigma_n^{-1}\left|\frac{\hat{h}_n'}{\hat{h}_n}\tilde{h}_n^{-1}(\tilde{h}_n - \hat{h}_n) + \tilde{h}_n^{-1}(\hat{h}_n' - \tilde{h}_n')\right|.\end{aligned} \tag{2.21}$$

Using (2.19)–(2.21) and (2.15) we get

$$\begin{aligned}&\int(\hat{q}_n' - \tilde{q}_n')^2\psi_n\tilde{h}_n(x)\,dx \\ &\quad= 0_p\Big(\sigma_n^{-4}\int(\tilde{h}_n - \hat{h}_n)^2\tilde{h}_n^{-1}\psi_n(x)\,dx + \int(\tilde{h}_n'' - \hat{h}_n'')^2\tilde{h}_n^{-1}\psi_n(x)\,dx \\ &\qquad\qquad + \sigma_n^{-2}\int(\hat{h}_n' - \tilde{h}_n')^2\tilde{h}_n^{-1}\psi_n(x)\,dx\Big) \\ &\quad= 0_p\left(\sigma_n^{-5}c_n^{-1}n^{-1}\right)\end{aligned}$$

by (2.18). From this, (2.16), and (2.18), we obtain (2.6). Similarly,

$$\begin{aligned}\int\bar{q}_n\psi_n\hat{h}_n'(x)\,dx - \hat{I}_n &= \frac{1}{4}\int\left(\hat{q}_n^2(x) - \hat{q}_n^2(-x)\right)\psi_n\hat{h}_n(x)\,dx \\ &= \frac{1}{4}\int\left(\hat{q}_n^2(x) - \hat{q}_n^2(-x)\right)\psi_n(\hat{h}_n - \tilde{h}_n)(x)\,dx \\ &= 0_p\left(\int|\hat{q}_n^2 - \tilde{q}_n^2|\psi_n|\hat{h}_n - \tilde{h}_n|(x)\,dx\right) \\ &= 0_p\Bigg(\left(\int|\hat{q}_n^2 - \tilde{q}_n^2|^2\psi_n\tilde{h}_n(x)\,dx\right)^{1/2} \\ &\qquad \times\left(\int(\hat{h}_n - \tilde{h}_n)^2\tilde{h}_n^{-1}\psi_n(x)\,dx\right)^{1/2}\Bigg) \\ &= 0_p\left(\sigma_n^{-3}c_n^{-1}n^{-1}\right).\end{aligned} \tag{2.22}$$

Proof of Lemma 2.3. For (2.8) write

$$\int(\bar{q}_n - \tilde{q}_n)\psi_n\hat{h}_n(x)\,dx = \int(\bar{q}_n - \tilde{q}_n)\psi_n(\hat{h}_n - \tilde{h}_n)(x)\,dx$$

and proceed as for (2.16).

For (2.9) write

$$\begin{aligned}\hat{I}_n - \int \tilde{q}_n^2 \psi_n \tilde{h}_n(x)\, dx &= \int \bar{q}_n^2 \psi_n (\hat{h}_n - \tilde{h}_n)(x)\, dx + \int (\bar{q}_n^2 - \tilde{q}_n^2) \psi_n \tilde{h}_n(x)\, dx \\ &= 0_p\left(\sigma_n^{-2} \left(\int (\hat{h}_n - \tilde{h}_n)^2 \tilde{h}_n^{-1} \psi_n(x)\, dx \right)^{1/2} \right. \\ &\qquad \left. + \sigma_n^{-1} \left(\int (\bar{q}_n - \tilde{q}_n)^2 \psi_n \tilde{h}_n(x)\, dx \right)^{1/2} \right) \\ &= 0_p\left(\sigma_n^{-5/2} c_n^{-1/2} n^{-1/2} \right)\end{aligned}$$

as in (2.21)–(2.22). □

Lemma 2.4 follows from Lemmas 2.1–2.3 and $\liminf_n I_n(\tilde{H}_n) > 0$, a consequence of Lemma 2.5 and our assumption on $\mathscr{H}$.

Proof of Lemma 2.5. For the proof of (2.11), without loss of generality suppose $\psi^{1/2}$ is continuously differentiable since for any ψ_1 satisfying our conditions and $\varepsilon > 0$, there exists a ψ_2 satisfying them such that $\psi_2^{1/2}$ is continuously differentiable and

$$(1 - \varepsilon)\psi_2(x) \le \psi_1(x) \qquad \text{for all } x.$$

If $H_n \xrightarrow{w} H_0$, $H_0(R) > 0$, and $\sup_n I(H_n) \le M < \infty$ by Cauchy–Schwarz,

$$\begin{aligned}|h_n^{1/2}(x) - h_n^{1/2}(y)| &= \frac{1}{2}\left| \int_x^y h_n^{-1/2} h_n'(t)\, dt \right| \\ &\le \frac{1}{2} I^{1/2}(H_n) |x - y|^{1/2} \\ &\le \frac{M^{1/2}}{2} |x - y|^{1/2}. \end{aligned} \tag{2.23}$$

Since $\int h_n(x)\, dx = 1$ for all n, (2.23) implies $\{h_n(x_0)\}$ bounded for any x_0. By Ascoli's theorem, (2.23) then implies $\{h_n^{1/2}\}$ and hence $\{h_n\}$ compact in the sup norm on $[-a, a]$ for all $a < \infty$. Since $H_n \xrightarrow{w} H_0$, a subsequence argument yields

$$h_n^{1/2}(x) \to h_0^{1/2}(x) \tag{2.24}$$

uniformly on $[-a, a]$. Next define an operator T_n on $L_2(R)$ by

$$T_n(v) = \frac{1}{2} \int_R h_n' h_n^{-1/2} \psi_n^{1/2} v(x)\, dx = \int_R \psi_n^{1/2} v(x)\, dh_n^{1/2}(x)$$

and

$$T(v) = \frac{1}{2}\int_R h_0' h_0^{-1/2} v(x)\, dx.$$

If v is continuously differentiable with compact support,

$$T_n(v) = -\int_R h_n^{1/2}\left(\left[\psi_n^{1/2}\right]' v + v'\psi_n^{1/2}\right)(x)\, dx \to -\int_R h_0^{1/2} v'(x)\, dx = T(v)$$

by (2.24) since the integrand is bounded and vanishes off a compact. Moreover

$$\begin{aligned} 4\|T_n\|^2 &= \int_R \psi_n \frac{[h_n']^2}{h_n}(x)\, dx \\ &= I_n(H_n) \le I(H_n) \le M. \end{aligned} \tag{2.25}$$

By the Banach Steinhaus theorem,

$$T_n(v) \to T(v) \tag{2.26}$$

for all v and

$$\liminf_n \|T_n\|^2 \ge \|T\|^2 = \frac{1}{4} I(H_0) \tag{2.27}$$

and (2.11) follows.

Since $\tilde{H}_n \xrightarrow{w} H_0$ and $I_n(\tilde{H}_n) \le I(\tilde{H}_n) \le I(H_n)$, (2.12) follows from (2.11).

Now take $\psi_n = 1$. The argument leading to (2.25)–(2.27) is valid. Therefore if $I(H_n) \to I(H_0)$, by (2.25) and (2.27),

$$\|T_n\| \to \|T\|. \tag{2.28}$$

But (2.26) and (2.28) imply

$$\|T_n - T\| \to 0$$

which is equivalent to (2.13).

Proof of Lemma 2.6. By Theorem 3.1, p. 124 of [7], we need only check that

$$\sup\left\{\int\left(h_n' h_n^{-1/2}\left(x - \frac{\theta}{n^{1/2}}\right) - h_n' h_n^{-1/2}(x)\right)^2 dx : |\theta| \le M\right\} \to 0,$$
$$\forall M < \infty,$$

and

$$\int\left[\frac{h_n'}{h_n}\right]^2 I\left[\left|\frac{h_n'}{h_n}\right| \ge \varepsilon n^{1/2}\right] h_n(x)\, dx \to 0, \qquad \forall \varepsilon > 0.$$

The first claim follows from (2.13) and the L_2 continuity theorem, the second from (2.13) and (2.24).

Proof of Theorem 1.1. By Lemmas 2.4 and 2.5, $\mathscr{L}_{H_n}(n^{1/2}\nu_n^*)$ and $\mathscr{L}_{H_n}(n^{-1/2}I^{-1}(H_0)\Sigma_{i=1}^n - \int\tilde{q}_n\psi_n(x)k_{\sigma_n}(x - X_i)\,dx)$ are asymptotically equal. Moreover,

$$\left|\int\tilde{q}_n\psi_n(x)k_{\sigma_n}(x - X_i)\,dx\right| \le \sigma_n^{-1} = o(n^{1/2})a.s. \tag{2.29}$$

and, by (2.12),

$$\begin{aligned}\limsup_n\int\left(\int\tilde{q}_n\psi_n(x)k_{\sigma_n}(x-z)\,dx\right)^2 h_n(z)\,dz \\ \le \limsup_n I_n(\tilde{H}_n) = I(H_0),\end{aligned}$$

if $H_n \xrightarrow{w} H_0$ and $I(H_n) \to I(H_0)$. By Lindeberg's theorem, the sequence $\mathscr{L}_{H_n}(n^{1/2}\nu_n^*)$ is then tight and all its limit points are $\mathscr{N}(0,\sigma^2)$ with $\sigma^2 \le I^{-1}(H_0)$. If $H_0(R) = 1$ and $I(H_0) < \infty$, by Lemma 2.6, and Cor. 11.1, p. 161 of [7], $\sigma^2 \ge I^{-1}(H_0)$ and the theorem follows. As a by-product we obtain

$$\int\left(\int\tilde{q}_n\psi_n(x)k_{\sigma_n}(x-z)\,dx\right)^2 h_n(z)\,dz \to I(H_0). \tag{2.30}$$

□

THEOREM 2.1. *Suppose $X_{1n},\dots,X_{nn}$ are independent, X_{in} has density $h_{in}(\cdot-\nu) = \theta_{in}^{-1}f(\theta_{in}^{-1}(\cdot-\nu))$, $i = 1,\dots,n$, f symmetric about* 0 *and $I(F) < \infty$. By H_n we denote the distribution function of $h_n = n^{-1}\Sigma_{i=1}^n h_{in}$. If H_n and H_0 satisfy the conditions of Theorem* 1.1, *then*

$$\mathscr{L}_{(0,\theta_{1n},\dots,\theta_{nn})}(n^{1/2}\nu_n^*) \to \mathscr{N}(0, I^{-1}(H_0)). \tag{2.31}$$

Proof of Theorem 2.1. The proofs of Lemmas 2.1–2.4 are essentially unchanged for this new model, as we can see by noting that the key inequalities (2.15) and (2.17) continue to hold. Moreover,

$$\begin{aligned}&\operatorname{var}_{(0,\theta_{ln},\dots,\theta_n)}\left(n^{-1/2}\sum_{i=1}^n\int\tilde{q}_n\psi_n(x)k_{\sigma_n}(x-X_i)\,dx\right) \\ &= \int\left(\int\tilde{q}_n\psi_n(x)k_{\sigma_n}(x-z)\,dx\right)^2 h_n(z)\,dz \to I(H_0),\end{aligned}$$

by (2.30). Consequently, (2.29) and Lindeberg's theorem yield (2.31).

NOTES. (1) If $f = \phi$, Theorem 2.1 shows that ν_n^* is regular (II′) and hence by Theorem 1.1 and Proposition 1.1 efficient (II′). We conjecture that it is in fact efficient within the class of all asymptotically normal translation equivariant estimates which are symmetric and even depend on G_n. That is, ν_n^* does as well as if we knew the θ_{in} up to a permutation.

(2) The companion problem, $X_i = (X_{i1}, X_{i2})$, X_{i1}, X_{i2} independent $\mathcal{N}(\theta_i, \nu)$ is much easier. Lindsay [10] and Hammerstrom [4] showed that the UMVU estimate $(2n)^{-1}\Sigma_{i=1}^n(X_{i1} - X_{i2})^2$ is efficient.

ACKNOWLEDGMENTS

We thank I. Johnstone, J. Ritov, and K. Takeuchi for helpful discussions.

REFERENCES

1. J. M. BEGUN, W. J. HALL, W. M. HUANG, AND J. A. WELLNER, Information and asymptotic efficiency in parametric–nonparametric models, *Ann. Statist.* **11** (1983), 432–452.
2. P. J. BICKEL, On adaptive estimation, *Ann. Statist.* **10** (1982), 647–671.
3. P. J. BICKEL, C. A. J. KLAASSEN, J. RITOV, AND J. A. WELLNER, "Efficient and Adaptive Statistical Inference," to be published by Johns Hopkins University Press, 1987.
4. I. J. GOOD, The population frequencies of species and the estimation of population parameters, *Biometrika* **40** (1953), 237–264.
5. T. HAMMERSTROM, "On Asymptotic Optimality Properties of Tests and Estimates in the Presence of Increasing Numbers of Nuisance Parameters," Ph.D. dissertation, University of California, Berkeley, 1978.
6. P. J. HUBER, Robust estimation of a location parameter, *Ann. Math. Statist.* **35** (1964), 73–101.
7. I. A. IBRAGIMOV AND R. Z. HASMINSKII, "Statistical Estimation: Asymptotic Theory," Springer-Verlag, New York, 1981.
8. J. KIEFER AND J. WOLFOWITZ, Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters, *Ann. Math. Statist.* **27** (1956), 887–906.
9. YU. A. KOSHEVNIK AND B. YA. LEVIT, On a non-parametric analogue of the information matrix, *Theory Probab. Appl.* **21** (1976), 738–753.
10. B. G. LINDSAY, Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators, *J. Philos. Trans. R. Soc. London Ser A* **296** (1980), 639–665.
11. B. G. LINDSAY, Efficiency of the conditional score in a mixture setting, *Ann. Statist.* **11** (1983), 486–497.
12. J. NEYMAN AND E. SCOTT, Consistent estimates based on partially consistent observations, *Econometrica* **16** (1948), 1–32.
13. J. PFANZAGL, "Contributions to a General Asymptotic Statistical Theory," Lecture Notes in Statistics, Springer Pub., New York, 1982.
14. H. ROBBINS, Asymptotically subminimax solutions of compound statistical decision problems, *in* "Proc. Second Berkeley Symp. Math. Statist. Probab.," pp. 131–148, Univ. of California Press, Berkeley, 1951.

15. H. ROBBINS, An empirical Bayes approach to statistics, *in* "Proc. Third Berkeley Symp. Math. Statist. Probab.," Vol. 1, pp. 157–163, Univ. of California Press, Berkeley, 1956.
16. H. ROBBINS, Some thoughts on empirical Bayes estimation, *Ann. Statist.* **11** (1983), 713–724.
17. C. STEIN, Efficient nonparametric testing and estimation, *in* "Proc. Third Berkeley Symp. Math. Statist. Prob.," Vol. 1, pp. 187–195, Univ. of California Press, Berkeley, 1956.
18. C. STONE, Adaptive maximum likelihood estimators of a location parameter, *Ann. Statist.* **3** (1975), 267–284.

The Annals of Statistics
1987, Vol. 15, No. 2, 513–540

EFFICIENT ESTIMATION IN THE ERRORS IN VARIABLES MODEL[1]

BY P. J. BICKEL AND Y. RITOV

University of California, Berkeley and The Hebrew University of Jerusalem

We consider efficient estimation of the slope in the errors in variables model with normal error when either the ratio of error variances is known and the distribution of the independent is arbitrary and unknown or the distribution of the independent variable is not Gaussian or degenerate. We calculate information bounds and exhibit estimates achieving these bounds using an initial minimum distance estimate and suitable estimates of the efficient score function.

1. Introduction. Errors in variables models have been the subject of an enormous amount of literature. A fairly recent reference with a good bibliography is Anderson (1984).

In its simplest form the model assumes n independent observations $\mathbf{X}_i = (X_i, Y_i)$, which are written as

$$X_i = X_i' + \varepsilon_{i1},$$
$$Y_i = \alpha + \beta X_i' + \varepsilon_{i2}. \tag{1.1}$$

The X_i' are viewed either as

(i) unknown constants;
(ii) independent identically distributed random variables.

Model (i) is called functional and (ii) structural by Kendall and Stuart (1979), Chapter 29.

The $(\varepsilon_{i1}, \varepsilon_{i2})$ are considered random vectors, which are identically distributed with mean 0, as well as independent of the X_i' in model (ii). In this paper we will deal exclusively with large sample theory in the structural model, although we believe our results generalize to the functional model. Our aim in this paper is the construction of efficient estimates of β under various assumptions in various special cases of (1.1). We also suggest how our results may be extended to instrumental variable models through the special case of repeated observations at the same X_i'.

Write $\mathbf{X}$, X', ε_1, ε_2 for "generic" observations. If we do not make any assumptions on the distributions of X and $(\varepsilon_1, \varepsilon_2)$, then β is clearly unidentifiable. In fact, β is unidentifiable even if we assume ε_1, ε_2 to be independent Gaussian variables with unknown variances and suppose X' is also Gaussian.

Received October 1984; revised September 1986.
[1]This research was supported in part by ONR Contract N00014-80-C-0163.
AMS 1980 *subject classifications*. Primary 62G20; secondary 62P20.
Key words and phrases. Reiersøl models, semiparametric, minimum distance, structural.

However, β has been shown to be identifiable under various sets of assumptions. These fall into two broad classes:

(A) *Gaussian errors.* $(\varepsilon_1, \varepsilon_2)$ have a bivariate Gaussian distribution with variance–covariance matrix Σ. The usual way to make β identifiable in the literature is to assume $\varepsilon_1, \varepsilon_2$ independent and either

$$\mathrm{Var}(\varepsilon_1) = c_0 \mathrm{Var}(\varepsilon_2) \tag{1.2}$$

or

$$\mathrm{Var}(\varepsilon_1) = c_0, \tag{1.3}$$

with c_0 assumed known. Both (1.2) and (1.3) are plausible under special circumstances [see Kendall and Stuart (1979), Chapter 29, for a discussion]. We shall explore a generalization of (1.2),

$$\Sigma = \sigma^2 \Sigma_0, \tag{1.4}$$

where Σ_0 is known. Model (1.3) can be analyzed in the same way. We shall call (1.4) the *restricted Gaussian error* model. This model and its generalizations to more complicated situations have been extensively studied; see Anderson (1984), for example. A second model in which the identifiability of β was established by Reiersøl (1950) puts no restriction on Σ but requires X' to be non-Gaussian (where constants are viewed as Gaussian). We shall call this the *general Gaussian error* model.

(B) *General independent errors.* Assume $\varepsilon_1, \varepsilon_2$ independent. If (1.2) holds, β is identifiable. This *restricted independent error* has also been extensively studied. If (1.2) is not present but either X' is non-Gaussian or $\varepsilon_1, \varepsilon_2$ have no Gaussian component, then, again according to Reiersøl (1950), β is identifiable. This *arbitrary independent error* model is probably most satisfactory but our results do not bear on it.

We review briefly some results on these models.

The restricted Gaussian model can be reduced to case (1.2) with $c_0 = 1$. The maximum likelihood estimate for β in this case is $\hat{\beta}_P$, which minimizes the sum of squared perpendicular distances of observed points from the fitted line

$$\sum_{i=1}^{n} \frac{(Y_i - \alpha - \beta X_i)^2}{1 + \beta^2}. \tag{1.5}$$

This estimate is well known to be $n^{1/2}$-consistent and asymptotically normal not only under the restricted Gaussian model but also under the restricted independent error model, see Gleser (1981) who considers multivariate generalizations. In the presence of fourth moments, it is not hard to show that $n^{1/2}$-consistency and asymptotic normality persist under the restricted independent error model when Σ_0 is the identity. Estimates of β in the general Gaussian error model, with Σ_0 diagonal, have been proposed by a variety of authors including Neyman and Scott (1948) and Rubin (1956). In the arbitrary independent error model, Wolfowitz in a series of papers ending in 1957, Kiefer and Wolfowitz (1956) and Spiegelman (1979) by a variety of methods gave estimates, which are consistent and in Spiegelman's case $n^{1/2}$-consistent and asymptotically normal.

Little seems to be known about the efficiency of these procedures other than that in the restricted Gaussian model the estimate $\hat{\beta}_P$ is efficient if X' is Gaussian by the classical results for M.L.E.'s in parametric models. Our main aims in this paper are:

In the general Gaussian error model:

(i) To give the structure that efficient estimates in the sense of Stein (1956), Koshevnik and Levit (1976) and Pfanzagl (1982) must have (Theorem 2.1).

(ii) To exhibit a reasonable efficient estimate (Theorem 2.2). In addition, we extend Theorem 2.1 to the simplest instrumental variable model, m repeated measurements with Gaussian errors,

$$X_{ij} = X_i' + \varepsilon_{ij1},$$
$$Y_{ij} = \alpha + \beta X_i + \varepsilon_{ij2}, \qquad j = 1,\dots,m,\ i = 1,\dots,r,\ n = mr,$$

and

$$\mathbf{X}_i = \{(X_{ij}, Y_{ij}),\ j = 1,\dots,m\},$$

where $m \geq 2$.

The ε_{ij2} are independent and identically distributed Gaussian and independent of ε_{ij1} which are also Gaussian. We refer this as the multiple *Gaussian measurements model*. Note that in this model if $m \geq 2$, the assumption of non-Gaussianity of the distribution of X' is unnecessary.

We speak of efficient estimation in the sense of Stein (1956) as developed by Koshevnik and Levit (1976), Pfanzagl (1982), Begun, Hall, Huang and Wellner (1983) and in a forthcoming monograph by Klaassen, Wellner and ourselves. Let $\mathbf{P}$ be the set of possible joint distributions of $\mathbf{X}$. We call $\mathbf{P}_0$ a parametric submodel of $\mathbf{P}$ if $\mathbf{P}_0 \subset \mathbf{P}$ and $\mathbf{P}_0$ can be represented as $\{P_{(\beta,\eta)};\ \beta \in \mathbf{R},\ \eta \in E$ open $\subset \mathbf{R}^k\}$. A parametric submodel is regular if at every (β_0, η_0) the mapping $(\beta, \eta) \to P_{(\beta,\eta)}$ is continuously Hellinger differentiable. Suppose that P belongs to $\mathbf{P}_0$—a regular parametric submodel of $\mathbf{P}$. Then the notion of information bound and efficient estimation of β are well defined [e.g., Ibragimov and Has'minskii (1981), pages 158–169]. Let $n^{-1}I^{-1}(P; \beta, \mathbf{P}_0)$ denote the asymptotic variance of an efficient estimate of β when P ranges over $\mathbf{P}_0$. Clearly, if we only assume that $P \in \mathbf{P}$ we can estimate no better than if we assumed that $P \in \mathbf{P}_0$. Accordingly, let $I(P; \beta, \mathbf{P}) = \inf\{I(P; \beta, \mathbf{P}_0)\colon \mathbf{P}_0$ a regular parametric submodel, $P \in \mathbf{P}_0\}$, be the information bound for estimating β under $\mathbf{P}$.

Loosely speaking, $\hat{\beta}_n$ is regular and efficient in $\mathbf{P}$ if

$$\mathbf{L}_P\big(\sqrt{n}\,(\hat{\beta}_n - \beta(P))\big) \to \mathbf{N}(0, I^{-1}(P; \beta, \mathbf{P})),$$

in some sense uniformly in $P \in \mathbf{P}$. Here $\mathbf{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. The weakest kind of uniformity acceptable is that

$$\mathbf{L}_{P_n}\big(\sqrt{n}\,(\hat{\beta}_N - \beta(P_n))\big) \to \mathbf{N}(0, I^{-1}(P; \beta, \mathbf{P})), \tag{1.6}$$

for sequences $P_n \in \mathbf{P}_0$, a regular parametric submodel as above, with $P_n = P_{(\beta_n, \eta_n)}$, $|\beta_n - \beta_0| = O(n^{-1/2}) = |\eta_n - \eta_0|$ for some β_0, η_0, $P = P_{(\beta_0, \eta_0)}$.

If $I^{-1}(P; \beta, \mathbf{P})$ is assumed at some $\mathbf{P}_0$, we obtain from the Hájek–Le Cam convolution theorem, Ibragimov and Has'minskii (1981), that $\hat{\beta}_n$ is asymptotically linear

$$\hat{\beta}_n = \beta(P) + n^{-1} \sum_{i=1}^{n} \tilde{l}(\mathbf{X}_i, P; \beta, \mathbf{P}) + o_p(n^{-1/2}),$$

where $\tilde{l}$ is defined as the efficient influence function, which has the properties

$$E_P \tilde{l}(\mathbf{X}_i, P; \beta, \mathbf{P}) = 0,$$
$$E_P \tilde{l}^2(\mathbf{X}_i, P, \beta, \mathbf{P}) = I^{-1}(P; \beta, \mathbf{P}).$$

Finding $\tilde{l}$ is equivalent to finding a suitable least favorable $\mathbf{P}_0$ (at each P). We discuss the theory which guides us in this search in Section 3.

Note that an estimate is efficient if

(a) it converges in law uniformly [as in (1.6)] on $\mathbf{P}$ and

(b) it is efficient in some parametric submodel $\mathbf{P}_0$ at each P. By the Hájek–Le Cam theorem (b) holds iff the efficient influence function is the influence function of the (local) maximum likelihood estimate of β in $\mathbf{P}_0$.

In Section 2 (Theorem 2.1), we exhibit $\tilde{l}$ and $\mathbf{P}_0$ for the general Gaussian error model and the restricted Gaussian model and discuss the main features of $I(P; \beta, \mathbf{P})$. In Theorem 2.2 we exhibit, for each of the two models, an estimate $\hat{\beta}$, converging in law uniformly [as in (1.6)] on $\mathbf{P}$, which has $\tilde{l}$ as influence function. By (a) and (b), $\hat{\beta}$ is necessarily efficient. The proof of Theorem 2.1 is deferred to Section 3, and the proof of Theorem 2.2 to Section 4.

2. The main results. Without loss of generality let $(\varepsilon_{i1}, \varepsilon_{i2}) \sim \mathbf{N}(0, \Sigma)$ where $\Sigma = [\sigma_{ij}]_{2\times 2}$ is nonsingular. Let $\theta = (\alpha, \beta, \Sigma)$ and

$$U(\theta) = U(\mathbf{X}, \theta) = \frac{Y - \alpha - \beta X}{\bar{\sigma}(\theta)}, \tag{2.1}$$

$$T(\theta) = T(\mathbf{X}, \theta) = \bar{\sigma}^{-2}(\theta)\big[(\sigma_{22} - \beta\sigma_{12})X + (\beta\sigma_{11} - \sigma_{12})(Y - \alpha)\big], \tag{2.2}$$

where $\bar{\sigma}^2(\theta)$ is the variance of $Y - \alpha - \beta X$ if θ is true,

$$\bar{\sigma}^2(\theta) = \beta^2\sigma_{11} - 2\beta\sigma_{12} + \sigma_{22}. \tag{2.3}$$

Then given θ, $T(\theta)$ is a complete and sufficient statistic for X' treated as a parameter, i.e., for the model $\{\mathbf{L}_\theta(\mathbf{X}|X' = \eta): \eta \in R\}$. This follows since given $X' = \eta$, (X, Y) have an $\mathbf{N}(\eta, \alpha + \beta\eta, \Sigma)$ distribution. Moreover, $U(\theta)$ is ancillary in this problem. It is necessarily independent of $T(\theta)$ in the original model and is distributed $\mathbf{N}(0,1)$. $T(\theta)$ is also the unbiased predictor of X', i.e., given $X' = \eta$, $T(\theta)$ has a $\mathbf{N}(\eta, \tilde{\sigma}^2(\theta))$ distribution, where

$$\tilde{\sigma}^2(\theta) = \bar{\sigma}^{-2}(\theta)\big(\sigma_{11}\sigma_{22} - \sigma_{12}^2\big).$$

We can write the joint density of $\mathbf{X}$ under (θ, G), where G is the distribution of X',

$$p(\mathbf{x}, \theta, G) = \int K(\mathbf{x}, z, \theta) G(dz), \tag{2.4}$$

where

$$\begin{aligned}K(\mathbf{x}, z, \theta) = {} & \left[2\pi\left(\sigma_{11}\sigma_{22} - \sigma_{12}^2\right)^{1/2}\right]^{-1} \\ & \times \exp\Big\{-\left[2\left(\sigma_{11}\sigma_{22} - \sigma_{12}^2\right)\right]^{-1} \\ & \qquad \times \left[\sigma_{22}(x - z)^2 - 2\sigma_{12}(x - z)\right. \\ & \qquad\qquad \left.\times (y - \alpha - \beta z) + \sigma_{11}(y - \alpha - \beta z)^2\right]\Big\} \\ = {} & \left[2\pi\left(\sigma_{11}\sigma_{22} - \sigma_{12}^2\right)^{1/2}\right]^{-1} \exp\left\{-\frac{1}{2}U^2(\mathbf{x}, \theta)\right\} \\ & \times \exp\left\{-\frac{\tilde{\sigma}^{-2}(\theta)}{2}(T(\mathbf{x}, \theta) - z)^2\right\},\end{aligned}$$

is the conditional density of $\mathbf{X}$ given $X' = z$.

Fix $\theta = \theta_0$, $G = G_0$. Drop the argument θ in $U(\theta)$, $T(\theta)$, $\bar{\sigma}^2(\theta)$, and $\tilde{\sigma}^2(\theta)$. Let

$$\omega(t) = \omega(t, \theta, G) = \tilde{\sigma}^{-1} \int \phi\left(\tilde{\sigma}^{-1}(t - z)\right) G(dz) \tag{2.5}$$

be the density of T and let

$$I_0 = \int \frac{[\omega']^2}{\omega}(t)\, dt$$

be the Fisher information for location of ω. Let $\eta = (\mu, \tau)$, $\mu \in R$, $\tau > 0$, and

$$G(\cdot, \eta) = G_0\left(\frac{\cdot - \mu}{\tau}\right).$$

Define

$$\mathbf{P}_0 = \left\{P_{(\theta, G(\cdot, \eta))}\right\}. \tag{2.6}$$

That is, in $\mathbf{P}_0$ we assume G known up to location and scale. $\mathbf{P}_0$ is not the same in the general Gaussian error model and the restricted Gaussian error model since Σ varies freely in the former!

THEOREM 2.1. *Assume $\int \eta^2 G(d\eta) < \infty$. Then $\mathbf{P}_0$ is the least favorable regular parametric submodel and the information bounds and the efficient influence functions for estimating β at $\theta = \theta_0$, $G = G_0$, are as follows*:

Restricted Gaussian error model. *Define the random variable*

$$l_a^* = \bar{\sigma}^{-1} U\left(T - E(T) + \tilde{\sigma}^2 \frac{\omega'}{\omega}(T)\right). \tag{2.7}$$

This is the efficient score function defined by Begun, Hall, Huang and Wellner (1983). *The information bound of* (1.5), *which we write as I_a, is given by*

$$\begin{aligned}I_a &= E_0(l_a^*)^2 = \bar{\sigma}^{-2}\left(\operatorname{Var}(T) + \tilde{\sigma}^4 I_0 - 2\tilde{\sigma}^2\right) \\ &= \bar{\sigma}^{-2}\left(\operatorname{Var}(X') + \tilde{\sigma}^2\left(\tilde{\sigma}^2 I_0 - 1\right)\right)\end{aligned} \tag{2.8}$$

and the efficient influence function is given by

$$\tilde{l}_a = l_a^*/I_a. \tag{2.9}$$

General Gaussian error model. *Define*

$$l_b^* = \tilde{\sigma}^{-1}U\left(T - E(T) + I_0^{-1}\frac{\omega'}{\omega}(T)\right). \tag{2.10}$$

The information bound is given by

$$\begin{aligned} I_b = E(l_b^*)^2 &= \tilde{\sigma}^{-2}\left(\mathrm{Var}(T) - I_0^{-1}\right) \\ &= \tilde{\sigma}^{-2}\left(\mathrm{Var}(X') + \tilde{\sigma}^2 - I_0^{-1}\right) \end{aligned} \tag{2.11}$$

and the efficient influence function by

$$\tilde{l}_b = l_b^*/I_b. \tag{2.12}$$

NOTES.

Restricted Gaussian error model.

(1) If $\sigma_{11} = 0$, then $\tilde{\sigma} = 0$ and we are in the case where $T = X = X'$ is observed without error. In this case,

$$I_a = \mathrm{Var}(X')/\mathrm{Var}(Y - \alpha - \beta X)$$

is the reciprocal of the asymptotic variance of $n^{1/2}$ times the ordinary least-squares estimate as it should be.

(2) If X' is normal, $\mathrm{Var}(T) = I_0^{-1}$ and (2.7) becomes

$$\begin{aligned} \tilde{\sigma}^{-2}\left(\mathrm{Var}(X') + \tilde{\sigma}^2(\tilde{\sigma}^2 - \mathrm{Var}(T))I_0\right) &= \tilde{\sigma}^{-2}(\mathrm{Var}\,X')\left(1 - \tilde{\sigma}^2 I_0\right) \\ &= \tilde{\sigma}^{-2}\mathrm{Var}^2(X')/\mathrm{Var}_0(T), \end{aligned}$$

which we shall call I_c.

This is just the asymptotic variance of $\hat{\beta}_P$ if $\Sigma_0 =$ identity [see, e.g., Gleser (1981)], whatever be G. So we conclude that we can do as well not knowing G as knowing it is Gaussian. This is a special instance of the claim that $\mathbf{P}_0$ given by (2.6) is least favorable.

(3) We can study the asymptotic efficiency I_c/I_a of $\hat{\beta}_P$ if G_0 is *not* normal. We show in Section 5 that, $I_c/I_a \geq (1 + \sigma^2/(\beta^2 + 1)(\mathrm{Var}(X') + \sigma^2))^{-1}$. In particular, if the signal-to-noise ratio in X, $\mathrm{Var}(X')/\sigma^2$, is large $\hat{\beta}_P$ is close to efficiency.

(4) The score function l_a^* can be written as

$$l_a^* = \tilde{\sigma}^{-1}U(E(X'|T) - E(X')).$$

The least-squares estimate if X' were known is based on the score function

$$\tilde{\sigma}^{-1}U(X' - E(X')).$$

Thus the efficient estimate replaces the unobservable X' by its best "estimate" $E(X'|T)$.

(5) Suppose that with $\Sigma = \sigma^2\Sigma_0$ we have m repeated observations at each X_i'. Then by sufficiency l_a^*, evaluated at the mean of each set of observations with Σ_0 replaced by Σ_0/m, is the efficient score function.

General Gaussian error model.

(1) Normality of X', under which β is unidentifiable, corresponds to G = point mass at 0. Appropriately, $I_b \to 0$ as G tends to point mass since then T approaches normality and $\tilde{\sigma}^2 \sim I_0^{-1}$.

(2) Necessarily, $I_a \geq I_b$. The inequality is always strict since

$$\tilde{\sigma}^2(I_a - I_b) = I_0^{-1}(\tilde{\sigma}^4 I_0^2 - 2\tilde{\sigma}^2 I_0 + 1)$$
$$= I_0^{-1}(\tilde{\sigma}^2 I_0 - 1)^2 > 0,$$

since I_0, the Fisher information for $X' + \varepsilon_1$, is always smaller than the Fisher information for ε_1 which is just $\tilde{\sigma}^{-2}$.

Multiple Gaussian measurements model. The efficient influence function can be calculated as for the general Gaussian error model, but is much more complicated.

Let $\mathbf{X} = (X_j, Y_j)$, $j = 1, \dots, m$, where $X_j = X' + \varepsilon_{j1}$, $Y_j = \alpha + \beta X' + \varepsilon_{j2}$ is a generic observation. We assume the ε_{ji} are independent Gaussian with mean 0 and $\operatorname{Var}(\varepsilon_{j1}) = \sigma_{11}$, $\operatorname{Var}(\varepsilon_{j2}) = \sigma_{22}$. Let

$$U = (\bar{Y} - \beta\bar{X} - \alpha)/\sigma_0, \tag{2.13}$$
$$T = (\sigma_{22}\bar{X} + \beta\sigma_{11}(\bar{Y} - \alpha))/(\sigma_{22} + \beta^2\sigma_{11}),$$

where $\bar{Y} = m^{-1}\Sigma_{j=1}^m Y_j$, $\bar{X} = m^{-1}\Sigma_{j=1}^m X_j$. Let

$$\sigma_0^2 = (\sigma_{22} + \beta^2\sigma_{11})/m,$$
$$\tilde{\sigma}^2 = \sigma_{11}\sigma_{22}/m^2\sigma_0^2, \tag{2.14}$$
$$I_0 = \int\left(\frac{w'}{w}\right)^2 w(t)\,dt, \quad \text{where } w \text{ is the density of } T \text{ given by (2.13).}$$

The efficient score function is then

$$l^* = \frac{UT}{\sigma_0\tilde{\sigma}^2} + a_2\frac{U}{\sigma_0\tilde{\sigma}^2}\frac{\omega'}{\omega}(T) + a_3(U^2 - 1) + a_4 S_1 + a_5 S_2, \tag{2.15}$$

where

$$S_1 = \sum_{j=1}^m \frac{(Y_j - \bar{Y})^2}{\sigma_{22}} - (m-1), \qquad S_2 = \sum_{j=1}^m \frac{(X_j - \bar{X})^2}{\sigma_{11}} - (m-1)$$

and the a's are functions of m, σ^2, σ_0^2 and I_0. For $m = 1$ the form of l^* agrees with l_b^* as it should. As $m \to \infty$,

$$a_2 \sim \tilde{\sigma}^2,$$

which corresponds to l_a^*. This is as expected since m large corresponds to σ_{11}, σ_{22} essentially known. The information I_d for this problem is I_b plus a complicated positive term vanishing for $m = 1$.

We now construct efficient estimates. The idea is to proceed as in the classical estimation of the location problem:

(a) Find a good estimate $\tilde{\beta}_n$ of β.

(b) (i) Consider $\tilde{l}$ as $\tilde{l}(\mathbf{x}, \beta, \eta, G)$ where $\theta = (\beta, \eta)$, G are now viewed as dummy variables and the argument $\mathbf{x}$ replaces $\mathbf{X}$. For example,

$$\tilde{l}_a(\mathbf{x}, \theta, G) = \tilde{\sigma}^{-1}(\theta) U(\mathbf{x}, \theta)\Big(T(\mathbf{x}, \theta) - \int T(\mathbf{x}, \theta) P_{(\theta, G)}(d\mathbf{x}) + \tilde{\sigma}^2(\theta) \frac{\omega'}{\omega} (T(\mathbf{x}, \theta), \theta)\Big) \Big/ I_a(\theta, G),$$

where T is given by (2.2) and $\omega(\cdot, \theta)$ is the marginal density of $T(\mathbf{X}, \theta)$, under $P_{(\theta, G)}$. Construct a suitable estimate $\hat{\tilde{l}}(\mathbf{x}, \beta; \mathbf{X}_1, \dots, \mathbf{X}_n)$ of $\tilde{l}(\mathbf{x}, \beta, \eta, G)$.

(ii) Form

$$\hat{\beta}_n = \tilde{\beta}_n + n^{-1} \sum_{i=1}^{n} \hat{\tilde{l}}(\mathbf{X}_i, \tilde{\beta}_n; \mathbf{X}_1, \dots, \mathbf{X}_n)$$

as the efficient estimate.

Preliminary estimate. We motivate our $\tilde{\beta}_n$ as follows. If we calculate under P_0 and $\beta = \beta_0$, $\mathrm{Var}(Y) \geq \mathrm{Var}(\beta X)$, then

$$\mathbf{L}(Y) = \mathbf{L}(\beta X + \sigma Z + \mu), \tag{2.16}$$

for $Z \sim \mathbf{N}(0,1)$ independent of X and

$$\mu = E(Y) - \beta E(X),$$
$$\sigma^2 = \mathrm{Var}(Y) - \beta^2 \mathrm{Var}(X).$$

If $\mathrm{Var}(Y) < \mathrm{Var}(\beta X)$, then

$$\mathbf{L}(X) = \mathbf{L}\left(\frac{Y}{\beta} + \sigma Z + \mu \right), \tag{2.17}$$

for $Z \sim \mathbf{N}(0,1)$ independent of Y, some σ, μ. For $|\beta| \neq |\beta_0|$ neither identity (2.16) nor (2.17) can hold; see Proposition 5.1. Our initial estimate is essentially a minimizing value for the distance between the natural estimates of the laws in (2.16) or (2.17). We believe our estimate may be improved by considering the joint distribution of (X, Y) and not only the marginals. For that note that if (2.16) holds, then

$$\mathbf{L}(\beta X + \sigma Z + \mu, Y) = \mathbf{L}(Y, \beta X + \sigma Z + \mu).$$

Another possible estimate is given by Spiegelman (1979) who does not assume Gaussianity of the errors but does assume ε_1, ε_2 independent. Different estimates $\tilde{\beta}_a$, $\tilde{\beta}_b$ are appropriate for the restricted Gaussian error model and the general Gaussian error model. Essentially, $\tilde{\beta}_b$ works whenever $\tilde{\beta}_a$ does except when G is

Gaussian. We give $\tilde{\beta}_b$ formally and sketch the difference for $\tilde{\beta}_a$. Without loss of generality, we assume $E(\varepsilon_1) = E(\varepsilon_2) = 0$.

Let $\hat{F}_1$ be the empirical distribution function of X_i, $i = 1, \dots, n$, and $F_1(\cdot)$ be the distribution function of X. Let $\hat{F}_2(\cdot)$ and $F_2(\cdot)$ be the empirical distribution function of Y_i and the distribution function of Y, respectively. Let

$$\hat{\mu}(\beta) = \bar{Y} - \beta\bar{X}, \qquad \hat{\sigma}^2(\beta) = |\hat{\sigma}_y^2 - \beta^2\hat{\sigma}_x^2|,$$

$$\hat{\sigma}_y^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \qquad \hat{\sigma}_x^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad \lambda = \hat{\sigma}_y/\hat{\sigma}_x. \tag{2.18}$$

Define, for $\hat{\sigma}_x^2 > 0$, $\hat{\sigma}_y^2 > 0$,

$$\begin{aligned}\Delta_n(\beta) &= \sqrt{n}\int\left|\hat{F}_2(y) - \int\Phi\left(\frac{y - \beta x - \hat{\mu}(\beta)}{\hat{\sigma}(\beta)}\right)d\hat{F}_1(x)\right|^2\phi(y)\,dy, \\ &\qquad\qquad \text{if } \hat{\sigma}_y^2 > \beta^2\hat{\sigma}_x^2 \\ &= \sqrt{n}\int\left|\hat{F}_1(x) - \int\Phi\left(\frac{\beta x - y + \hat{\mu}(\beta)}{(\operatorname{sgn}\beta)\hat{\sigma}(\beta)}\right)d\hat{F}_2(y)\right|^2\lambda\phi(\lambda y)\,dy, \\ &\qquad\qquad \text{if } \hat{\sigma}_y^2 < \beta^2\hat{\sigma}_x^2.\end{aligned} \tag{2.19}$$

Note that $\Delta_n(\beta)$ can be defined by continuity at $\sigma(\beta) = 0$ since $P[|\beta| + \hat{\sigma}^2(\beta) > 0, \forall\beta] = 1$. For given $a > 0$, let $\Delta_n(\beta, a)$ be the corresponding quantity with Y_i replaced by $Y_i + aX_i$, $i = 1, \dots, n$. Let $\beta_n^*(a)$ minimize $\Delta_n(\beta, a)$. $\beta_0 = 0$ poses difficulties but we can always shift away from this value. Accordingly, let

$$\begin{aligned}\beta_n^* &= \beta_n^*(0), &&\text{if } |\beta_n^*(0)| \ge \delta_0 \\ &= \beta_n^*(2\delta_0) - 2\delta_0, &&\text{if } |\beta_n^*(0)| < \delta_0.\end{aligned}$$

Finally, we need to distinguish between $\pm\beta_n^*$. For that let $\hat{W}_n^+$ be the empirical distribution function of $\hat{\sigma}^{-1}(Y_i - \mu(\beta_n^*) - \beta_n^* X_i)$, where

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \mu(\beta_n^*) - \beta_n^* X_i)^2$$

and $\hat{W}_n^-$ the corresponding quantity for $-\beta_n^*$. Let

$$\begin{aligned}\tilde{\beta} &= \beta_n^*, \quad \text{if } \int|\hat{W}_n^+(y) - \Phi(y)|^2\phi(y)\,dy \le \int|W_n^-(y) - \Phi(y)|^2\phi(y)\,dy \\ &= -\beta_n^*, \quad \text{otherwise.}\end{aligned}$$

For the restricted Gaussian error model, $\Sigma_0 =$ identity we proceed as above but change the definition of $\hat{\sigma}^2(\beta)$ to, using the new information,

$$\hat{\sigma}_a^2(\beta) = \frac{|1 - \beta^2|}{1 + \beta^2}n^{-1}\sum_{i=1}^{n}(Y_i - \hat{\mu}(\beta) - \beta X_i)^2$$

and switch the definition of $\Delta_n(\beta)$ as $\beta^2 \le 1$ or > 1.

EFFICIENT ESTIMATES. Note that

$$\beta\sigma_{11} - \sigma_{12} = \beta \operatorname{Var}(X) - \operatorname{cov}(X, Y), \tag{2.20}$$

$$\sigma_{22} - \beta\sigma_{12} = \operatorname{Var}(Y) - \beta \operatorname{cov}(X, Y), \tag{2.21}$$

$$\alpha = E(Y) - \beta E(X). \tag{2.22}$$

We can reparametrize the general Gaussian error model using $(\beta, \alpha, \gamma_1, \gamma_2, \sigma_{11}, G)$, where $\gamma_1, \gamma_2, \alpha$ are the expressions in (2.20)–(2.22), respectively. Abusing notation, let $\theta = (\beta, \alpha, \gamma_1, \gamma_2)$ so that

$$U_i(\theta) = (Y_i - \alpha - \beta X_i)/(\beta\gamma_1 + \gamma_2)^{1/2},$$

$$T_i(\theta) = (\gamma_2 X_i + \gamma_1(Y_i - \alpha))/(\beta\gamma_1 + \gamma_2).$$

Define $\tilde{\theta}_n = (\tilde{\beta}_n, \tilde{\alpha}_n, \tilde{\gamma}_{1n}, \tilde{\gamma}_{2n})$ by substituting sample moments and $\tilde{\beta}_n$ in the definitions (2.20)–(2.22) for $\beta, \alpha, \gamma_1, \gamma_2$. Let

$$\lambda(t) = e^{-t}(1 + e^{-t})^2,$$

$$\lambda_\nu(t) = \frac{1}{\nu}\lambda\left(\frac{t}{\nu}\right).$$

For sequences c_n, $\nu_n \downarrow 0$, to be characterized later, let $\lambda_n = \lambda_{\nu_n}$ and estimate ω_0 by the kernel estimator,

$$\hat{\omega}_n(t, \theta) = \frac{1}{n}\sum_{i=1}^{n} \lambda_\nu(t - T_i(\theta)) + c_n.$$

Define the efficient estimate for the general Gaussian error model by

$$\tilde{\beta}_{nb} = \tilde{\beta}_n + n^{-1}\hat{I}_b^{-1}\sum_{i=1}^{n}\frac{\tilde{U}_i}{\sigma(\tilde{\theta}_n)}\left(\tilde{T}_i - \tilde{T}. + \hat{I}_0^{-1}\frac{\hat{\omega}_n'}{\hat{\omega}_n}(\tilde{T}_i, \bar{\theta}_n)\right), \tag{2.23}$$

where $\tilde{U}_i$, $\tilde{T}_i$ are used for $U_i(\tilde{\theta}_n)$, $T_i(\tilde{\theta}_n)$, and $\tilde{T}. = n^{-1}\Sigma_{i=1}^n \tilde{T}_i$,

$$\hat{I}_0 = n^{-1}\sum_{i=1}^{n}\left(\frac{\hat{\omega}_n'}{\hat{\omega}_n}\right)^2(T_i(\tilde{\theta}_n), \tilde{\theta}_n), \tag{2.24}$$

$$\hat{I}_b = (\tilde{\beta}_n\tilde{\gamma}_{12} + \tilde{\gamma}_{2n})^{-1}n^{-1}\sum_{i=1}^{n}\left(\tilde{T}_i - \tilde{T}. + \hat{I}_0^{-1}\frac{\omega_n'}{\omega_n}(\tilde{T}_i, \tilde{\theta}_n)\right)^2. \tag{2.25}$$

Similarly, we define the efficient estimate $\hat{\beta}_{na}$ for the restricted Gaussian error model by

$$\hat{\beta}_{na} = \tilde{\beta}_{na} + n^{-1}\hat{I}_a^{-1}\sum_{i=1}^{n}\frac{\tilde{U}_i}{\hat{\sigma}_n}\left(\tilde{T}_{ia} - \tilde{T}._a + (1 + \tilde{\beta}_n^2)^{-1}\hat{\sigma}_n^2\frac{\hat{\omega}_n'}{\hat{\omega}_n}(\tilde{T}_{ia}, \tilde{\theta}_n)\right),$$

where

$$\hat{\sigma}_n^2 = \left(1 + \tilde{\beta}_n^2\right)^{-1} n^{-1} \sum_{i=1}^{n} \left(Y_i - \mu(\tilde{\beta}_n) - \tilde{\beta}_n X_i\right)^2,$$

$$\tilde{T}_{ia} = \left(\tilde{\beta}_n(Y_i - \tilde{\alpha}_n) + X_i\right)\left(1 + \tilde{\beta}_n^2\right)^{-1},$$

$$\hat{I}_a = \hat{\sigma}_n^{-2} n^{-1} \sum_{i=1}^{n} \left(\tilde{T}_{ia} - \tilde{T}_{\cdot a} + \left(1 + \tilde{\beta}_n^2\right)^{-1} \hat{\sigma}_n^2 \frac{\omega_n'}{\omega_n}\left(\tilde{T}_{ia}, \tilde{\theta}_n\right)\right)^2,$$

in accordance with (2.7) and (2.8).

Let $\{c_n\}, \{\nu_n\}$ be such that

$$c_n \to 0, \qquad \nu_n \to 0, \qquad n c_n^2 \nu_n^6 \to \infty.$$

THEOREM 2.2. (i) *Suppose G_0 is non-Gaussian, $\int x^2\, dG_0(x) < \infty$ and $\mathbf{P}_0 = \{P_{(\theta, G_0)}: \theta \in \Theta\}$ is regular. Then, if $P_0 = P_{(\theta_0, G_0)}$ satisfies the general Gaussian error model,*

$$\mathbf{L}_{P_0}\left(n^{1/2}\left(\hat{\beta}_{bn} - \beta(P_0)\right)\right) \to \mathbf{N}\left(0, I_b^{-1}(P_0)\right), \tag{2.26}$$

for all $P_0 \in \mathbf{P}_0$.

(ii) *If also $n\nu_n^{-6}\log n \to 0$, the convergence in* (2.26) *continues to hold if P_0 is replaced by $P_n = P_{(\theta_n, G_n)}$, where*

$$\theta_n = (\beta_n, \alpha_n, \gamma_{1n}, \gamma_{2n}, \sigma_{11n}) \to \theta = (\beta, \alpha, \gamma_1, \gamma_2, \sigma_{11})$$

and $G_n \to G$ weakly and $\int z^2 G_n(dz) \to \int z^2 G(dz) < \infty$.

(iii) *Write* (2.21)–(2.23) *as $\hat{\beta}_n = \hat{\beta}_n(\tilde{\beta}_n)$ and let $\hat{\beta}_{0n} = \tilde{\beta}_n$, $\hat{\beta}_{in} = \hat{\beta}_n(\hat{\beta}_{i-1,n})$, $i = 1, 2, 3, \ldots$. Then, for $i \geq 1$, all $\hat{\beta}_{in}$ are efficient and $|\hat{\beta}_{in} - \hat{\beta}_{i-1,n}| = o_p(n^{-1/2})$ for all $i \geq 2$.*

(iv) *If $\hat{\beta}_n$ is replaced by $\hat{\beta}_{an}$ and the restricted Gaussian error model is considered then claims* (i)–(iii) *continue to hold with I_b replaced by I_a.*

NOTES.

(1) Let $K \subset \mathbf{P}$ be compact in the total variation norm topology. Part (ii) of the theorem shows that the convergence in (2.24) is uniform over K if $P \to I_b(P)$ is continuous on K. These are the largest sets over which we may expect uniform convergence.

(2) Part (iii) of the theorem may be interpreted in terms of running the iteration $\hat{\beta}_{in}$ to convergence. Suppose the stopping rule is of the form: Stop as soon as $|\hat{\beta}_{in} - \beta_{i-1,n}| \leq \varepsilon_n$, where $\varepsilon_n \downarrow 0$, $n^{1/2}\varepsilon_n > c > 0$. This is reasonable since the random fluctuations in the estimate are of order $n^{-1/2}$. Then, by part (iii), with probability tending to 1 the iteration stops with $\hat{\beta}_{2n}$.

Under more stringent conditions on ν_n, c_n we conjecture that tedious calculations will show that, in fact, $\lim_i \hat{\beta}_{in}$ exists with probability tending to 1 and is efficient.

3. Information bounds and proof of Theorem 2.1. Let $\mathbf{P}_0$ be a regular parametric submodel of a model $\mathbf{P}$ written in the form $\{P_{(\beta,\gamma)}: \beta \in R, \gamma \in E \subset R^k\}$. Let $l(X, \beta, \gamma)$ denote the log likelihood of an observation from $P_{(\beta,\gamma)}$ and let $\dot{l}_0(X) = \partial l/\partial\beta|_{(\beta_0,\gamma_0)}$, $\dot{l}_j(X) = \partial l/\partial\gamma_j|_{(\beta_0,\gamma_0)}$, $1 \le j \le k$, where $\gamma = (\gamma_1, \dots, \gamma_k)$. Begun, Hall, Huang and Wellner (1983) [see also Efron (1977) and Neyman (1957)] show (in slightly different terms) that, if $P_0 = P_{(\beta_0,\gamma_0)}$

$$I(P_0; \beta, \mathbf{P}_0) = \min\left\{E\left(\dot{l}_0(X) - \sum_{j=1}^k c_j \dot{l}_j(X)\right)^2 : (c_1, \dots, c_k) \in R^k\right\}$$
$$= E\{[l^*]^2(X)\},$$

where

$$l^* = \dot{l}_0 - \sum_{j=1}^k c_j^* \dot{l}_j, \tag{3.1}$$

and the c_j^* are uniquely determined by the orthogonality condition

$$El^*\dot{l}_j(X) = 0, \qquad j = 1, \dots, k. \tag{3.2}$$

Moreover, the efficient influence function for $\mathbf{P}_0$ is given by

$$\tilde{l}(X, P_0|\beta, \mathbf{P}_0) = l^*(X)/I(P_0; \beta, \mathbf{P}_0). \tag{3.3}$$

Therefore, to calculate $\tilde{l}$ for $\mathbf{P}_0$ we need only calculate the projection $\sum_{j=1}^k c_j^* \dot{l}_j(X)$, in $L_2(P_0)$, of $\dot{l}_0$ into $[\dot{l}_j: 1 \le j \le k]$, the linear span of $\dot{l}_1, \dots, \dot{l}_k$. Let $\Pi(h|L)$ denote the projection of $h \in L_2(P_0)$ into a closed linear space $L \subset L_2(P_0)$.

To prove Theorem 2.1 we go through the following steps for the restricted Gaussian error model and an analogous series for the general Gaussian error model.

(i) Identify $(\gamma_1, \gamma_2) = (\alpha, \sigma^2)$, where σ^2 is given by (1.4) and let $\eta = (\eta_1, \dots, \eta_{k-2})$ index G, i.e.,

$$\mathbf{P}_0 = \left\{P_{(\theta, G_\eta)}: \eta \in E, \theta = (\alpha, \beta, \sigma^2), \alpha, \beta \in R\right\}.$$

Calculate formally $\dot{l}_j$, $0 \le j \le k$, at $P_0 = P_{(\theta_0, G_{\eta_0})}$, where $j = 0 \leftrightarrow \beta$, $j = 1, 2 \leftrightarrow \alpha, \sigma^2$, $j \ge 3 \leftrightarrow \eta$.

We project $\dot{l}_0$ into $[\dot{l}_j: j \ge 1]$ in two steps. First, calculate, for $0 \le j \le 2$, $\Pi(\dot{l}_j|V)$, where

$$\begin{aligned} V &= [\dot{l}_j: j \ge 3], \\ l^* &= \dot{l}_0 - \Pi(\dot{l}_0|V) - \Pi(\dot{l}_0 - \Pi(\dot{l}_0|V)|W), \end{aligned} \tag{3.4}$$

where

$$W = [\dot{l}_j - \Pi(\dot{l}_j|V): 1 \le j \le 2].$$

Claim (3.4) is well known and can be verified by checking (3.2). We establish that:

(ii) For any regular parametric submodel $\mathbf{P}_0$

$$[\dot{l}_j\colon j \geq 3] \subset \{a(T)\colon a(T) \in L_2(P_0),\ Ea(t) = 0\}$$

and then prove:

(iii) If $\mathbf{P}_0$ is given by (2.6), then $\mathbf{P}_0$ is regular and

$$[\dot{l}_j\colon j \geq 3] \supset [E(\dot{l}_0(X)|T)]. \tag{3.5}$$

The existence of a model $\mathbf{P}_0$ having property (3.5), but not the specific choice (2.6), follows from Theorem 14.3.12 of Pfanzagl (1982). Note that

$$E(h(X) - E(h(X)|T))a(T) = 0, \quad \text{for all } a(T),\ h \in L_2(P_0). \tag{3.6}$$

Now (ii) and (iii) imply that, for $\mathbf{P}_0$ given by (2.6),

$$\Pi(\dot{l}_i|V) = E(\dot{l}_i(X)|T), \qquad 0 \leq i \leq 2,$$

and hence by (3.4) if l_0^* is the l^* of $\mathbf{P}_0$ given by (2.6),

$$l_0^*(X) = \dot{l}_0(X) - E(\dot{l}_0(X)|T) - \sum_{j=1}^{2} d_j(\dot{l}_j(X)) - E(\dot{l}_j(X)|T), \tag{3.7}$$

with $\{d_j\colon 1 \leq j \leq 2\}$ determined by (3.2) for $j = 1, 2$. Take $\mathbf{P}_0$ to be any regular parametric submodel. By (ii) and (3.6)

$$El_0^*(X)\dot{l}_j(X) = 0, \qquad j \geq 3.$$

By (3.2)

$$El_0^*(X)\dot{l}_j(X) = 0, \qquad j = 1, 2.$$

Therefore,

$$\begin{aligned} &E(l^*(X))^2 - E(l_0^*(X))^2 \\ &\quad = E(l^*(X) - l_0^*(X))^2 + 2E(l_0^*(X)(l^* - l_0^*)(X)) \\ &\quad = E(l^*(X) - l_0^*(X))^2 \geq 0, \end{aligned} \tag{3.8}$$

since $l^* - l_0^* \in [\dot{l}_j\colon j \geq 1]$. We conclude that $\mathbf{P}_0$ given by (2.6) is least favorable.

Proof of Theorem 2.1. For mnemonic convenience we write $\dot{l}_0 = l_\beta$ and $\dot{l}_j = l_\alpha,\ l_{\sigma^2},\ l_{\sigma_{11}}$, etc., as appropriate.

Restricted Gaussian error model. (i) Differentiating (2.4) we get, for $\theta = \theta_0$, $G = G_0$,

$$\begin{aligned} l_\beta(\mathbf{X}) &= p^{-1}(\mathbf{X}, \theta, G)\int(\sigma_{11}\sigma_{22} - \sigma_{12}^2)^{-1}(\sigma_{11}(Y - \alpha - \beta z) - \sigma_{12}(X - z)) \\ &\qquad \times zK(\mathbf{X}, z, \theta)G(dz) \\ &= \tilde{\sigma}^{-1}(\theta)\int\left(\frac{U}{\sigma(\theta)} + \frac{\beta\sigma_{11} - \sigma_{12}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}(T - z)\right) \\ &\qquad \times z\phi(\tilde{\sigma}^{-1}(\theta)(T - z))G(dz)/\omega(T), \end{aligned} \tag{3.9}$$

since

$$X = T - \tilde{\sigma}^{-1}(\theta)(\beta\sigma_{11} - \sigma_{12})U,$$
$$Y - \alpha = \beta T + \tilde{\sigma}^{-1}(\theta)(\sigma_{22} - \beta\sigma_{12})U.$$

Similarly,

$$(3.10)\qquad \begin{aligned} l_\alpha &= [\omega(T)\tilde{\sigma}(\theta)]^{-1} \\ &\quad\times\int\left(\frac{U}{\sigma(\theta)} + \frac{\beta\sigma_{11} - \sigma_{22}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}(T - z)\right)\phi(\tilde{\sigma}^{-1}(\theta)(T - z))G(dz), \end{aligned}$$

$$(3.11)\qquad \begin{aligned} l_{\sigma^2} &= \frac{1}{2\sigma^2}\Big((U^2 - 1) + \tilde{\sigma}^{-1}(\theta) \\ &\quad\times\int\left(\tilde{\sigma}^{-2}(\theta)(T - z)^2 - 1\right)\phi(\tilde{\sigma}^{-1}(\theta)(T - z))G(dz)/\omega(T)\Big). \end{aligned}$$

(ii) Suppose $\mathbf{P}_0 = \{P_{(\theta, G_\eta)}\}$ is a regular submodel with $G_\eta \ll G_0 = G$. If $g_\eta = dG_\eta/dG$, $g_0 = 1$, and, formally,

$$(3.12)\qquad \dot{l}_{j+2}(X) = \int\exp\left\{-\frac{\tilde{\sigma}^{-2}}{2}(T - z)^2\right\}\frac{\partial g_\eta}{\partial\eta_j}(z)G(dz)/\omega(T),$$

a function of T only. If $\dot{l}_{j+2}$ exists only in the Hellinger sense it is easy to check that $\dot{l}_{j+2}$ is an L_2 limit of functions of T and hence T measurable.

(iii) If $\mathbf{P}_0$ is given by (2.6),

$$(3.13)\qquad \begin{aligned} \frac{\partial l}{\partial\mu}(X,\theta,G_\eta)\bigg|_{\mu=0,\,\tau=1} &= \omega^{-1}(T)\frac{\partial}{\partial\mu}\int\exp\left\{-\frac{\tilde{\sigma}^{-2}}{2}\left(T - \frac{(z-\mu)}{\tau}\right)^2\right\}G(dz) \\ &= \omega^{-1}(T)\int(T - z)\exp\left\{-\frac{\tilde{\sigma}^{-2}}{2}(T - z)^2\right\}G(dz), \end{aligned}$$

$$(3.14)\qquad \frac{\partial l}{\partial\tau}(X,\theta,G_\eta)\bigg|_{\mu=0,\,\tau=1} = \omega^{-1}(T)\int z(T - z)\exp\left\{-\tilde{\sigma}^{-2}(T - z)^2\right\}G(dz).$$

The independence of U and T and $EU = 0$ yield from (3.9)

$$E(l_\beta|T) = \tilde{\sigma}^{-1}\frac{\beta\sigma_{11} - \sigma_{12}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\int z(T - z)\phi(\tilde{\sigma}^{-1}(T - z))G(dz)/\omega_0(T),$$

which is proportional to $\partial l/\partial\tau$ as required. Therefore,

$$l_\beta - E(l_\beta|T) = \tilde{\sigma}^{-1}U\left[\int\phi(\tilde{\sigma}^{-1}(T - z))G(dz)\right]^{-1}\int z\phi(\tilde{\sigma}^{-1}(T - z))G(dz).$$

From (2.5)

$$(3.15)\qquad l_\beta - E(l_\beta|T) = \tilde{\sigma}^{-1}U\left(T + \tilde{\sigma}^2\frac{\omega'}{\omega}(T)\right).$$

Similarly,

$$l_\alpha - E(L_\alpha|T) = \tilde{\sigma}^{-1}U,$$

$$l_{\sigma^2} - E(l_{\sigma^2}|T) = \frac{1}{2\tilde{\sigma}^2}(U^2 - 1).$$

Now, from (3.10) and (3.13)

$$l_\alpha - \Pi(l_\alpha|V) = l_\alpha - E(l_\alpha|T) \tag{3.16}$$

and necessarily by (ii)

$$l_{\sigma^2} - \Pi(l_{\sigma^2}|V) = l_{\sigma^2} - E(l_{\sigma^2}|T) + b(T). \tag{3.17}$$

Therefore, (3.17) is orthogonal to both (3.16) and (3.15) so that $d_2 = 0$. On the other hand, it is easy to see that $d_1 = E(T)$. From (3.7), (3.15) and (3.16) we obtain Theorem 2.1 for a restricted Gaussian model.

General Gaussian error model. We find after some computation

$$\begin{aligned} l_{\sigma_{11}} &= \alpha_{11}(U^2 - 1) + \beta_{11}U\frac{\omega'}{\omega}(T) + \gamma_{11}b(T), \\ l_{\sigma_{22}} &= \alpha_{22}(U^2 - 1) + \beta_{22}U\frac{\omega'}{\omega}(T) + \gamma_{22}b(T), \\ l_{\sigma_{12}} &= \alpha_{12}(U^2 - 1) + \beta_{12}U\frac{\omega'}{\omega}(T) + \gamma_{12}b(T), \end{aligned} \tag{3.18}$$

where

$$b(T) = \tilde{\sigma}^{-1}\int z^2\phi(\tilde{\sigma}^{-1}(T - z))G(dz)/\omega(T),$$

and the matrix $\begin{pmatrix} \alpha_{11} & \beta_{11} \\ \alpha_{22} & \beta_{22} \\ \alpha_{12} & \beta_{12} \end{pmatrix}$ has dimension 2. Let $V = [l_\mu(\mathbf{x}), l_\tau(\mathbf{x})]$.

From (3.18) the linear span of $l_\alpha - E(l_\alpha|T)$, $l_{\sigma_{ij}} - \Pi(l_{\sigma_{ij}}|V)$, $i, j = 1, 2$, is

$$\left[U, U^2 - 1, U\frac{\omega'}{\omega}(T), c(T)\right], \tag{3.19}$$

where $c(T) = \Pi(b(T)|V)$. We find the projection of $l_\beta - E(l_\beta|T)$ on (3.19) by using the independence of U and T, $EU = 0$, $EU^2 = 1$. We obtain

$$\begin{aligned} &\tilde{\sigma}^{-1}(\Pi(UT|[U])) + \Pi\left(UT\middle|\left[U\frac{\omega'}{\omega}(T)\right]\right) + \tilde{\sigma}^2U\frac{\omega'}{\omega}(T) \\ &= \tilde{\sigma}^{-1}UE(T) + \left(\tilde{\sigma}^2 - \frac{1}{I_0}\right)U\frac{\omega'}{\omega}(T), \end{aligned}$$

since $E(T(\omega'/\omega)(T) = -1$. We conclude that under the submodel (2.6), with Σ varying freely, l_b^* is the efficient score function. But clearly, $El_b^*(\mathbf{X})a(T) = 0$ for all $a(T) \in L_2(P_0)$ and, in view of (ii), the argument leading to (3.8) applies to l_b^* also and (2.6) is least favorable. □

4. Proof of Theorem 2.2 and miscellaneous results. We begin by studying β_n^*.

PROPOSITION 4.1. *If either*

$$\mathbf{L}_{P_0}(Y) = \mathbf{L}_{P_0}(\beta X) * \mathbf{N} \tag{4.1}$$

or

$$\mathbf{L}_{P_0}(X) = \mathbf{L}_{P_0}(Y/\beta) * \mathbf{N} \tag{4.2}$$

(*where* $\mathbf{N}$ *is a Gaussian law and* $*$ *denotes convolution*), *then* $|\beta| = |\beta_0|$ *or* G_0 *is Gaussian. If* $\beta = \beta_0$ *one of these relations holds.*

PROOF. Let ψ be the characteristic function of X'. The case $\beta_0 = 0$ is simple. Assume $\beta_0 \neq 0$. Without loss of generality, take $E_0(X) = E_0(Y) = 0$ and $\beta_0 = 1$. Suppose $|\beta| \neq 1$ and without loss of generality, take $|\beta| > 1$. Then (4.1) becomes

$$\psi(t) = \psi(\beta t) e^{at^2}, \tag{4.3}$$

for some a. Iterating (4.3) we get for all k, t

$$\psi(\beta^k t) = \exp\left(-at^2 \frac{(\beta^{2k} - 1)}{(\beta^2 - 1)}\right) \psi(t).$$

Putting $u = \beta^k t$ and letting $k \to \infty$,

$$\psi(u) = \exp\left(-au^2(\beta^2 - 1)^{-1}(1 + o(1))\right)(1 + o(1))$$

and we get G_0 Gaussian. The same argument works for (4.2). $\square$

PROPOSITION 4.2. *Suppose that* $\mathbf{P}$ *consists of all probabilities satisfying the general Gaussian error model with* $\int x^2\, dG(x) < \infty$. *Then for every* $P_0 \in \mathbf{P}$

$$\lim_{M \to \infty} \lim_{n \to \infty} P_0\left[\sqrt{n}\, \tilde{\beta}_n - \beta(P_0)| \geq M\right] = 0.$$

PROOF. Let

$$\begin{aligned} Z_n(y, \beta) &= \sqrt{n}\left\{(\hat{F}_2(y) - F_2(y)) - \int \Phi\left(\frac{y - \beta x - \mu(\beta)}{\sigma(\beta)}\right) d(\hat{F}_1(x) - F_1(x))\right\} \\ &= \sqrt{n}\left\{(\hat{F}_2(y) - F_2(y)) + \operatorname{sgn}\beta \int \left(\hat{F}_1((y - \mu(\beta) - z\sigma(\beta))/\beta)\right.\right. \\ &\qquad \left.\left. - F_1((y - \mu(\beta) - z\sigma(\beta))/\beta)\right)\phi(z)\, dz\right\}, \end{aligned} \tag{4.4}$$

where F_1, F_2 are the marginal distribution functions of X and Y under P_0 and $\mu(\beta)$, $\sigma(\beta)$ are obtained by substituting population for sample moments in (2.18) and (2.19). By strong approximation, e.g., Csörgő (1981), we can construct $Z(\cdot,\cdot)$, a mean 0 Gaussian process in $C([-\infty, \infty] \times [-\infty, \infty])$ such that

$$\sup_{y, \beta} |Z_n(y, \beta) - Z(y, \beta)| = o_p(1). \tag{4.5}$$

Let $\hat{Z}_n(\cdot,\cdot)$ be defined by replacing $\mu(\beta), \sigma(\beta)$ by $\hat{\mu}(\beta), \hat{\sigma}(\beta)$ in (4.4). For $\sigma(\beta) \geq \varepsilon$, the family of functions $x \to \Phi((y - \beta x - \mu(\beta))/\sigma(\beta))$ is uniformly bounded and equicontinuous. Moreover,

$$\sup\big\{\sigma^{-1}(\beta)\beta\phi\big((y - \beta x - \mu(\beta)/\sigma(\beta)\big) \\ -\hat{\sigma}(\beta)\beta\phi\big(y - \beta x - \mu(\hat{\beta})\big)/\sigma(\hat{\beta})\big) \colon \sigma(\beta) \geq \varepsilon\big\} \to_P 0.$$

From (4.4) we then conclude that

$$\sup_y \big\{ |\hat{Z}_n(y,\beta) - Z_n(y,\beta)| \colon \sigma(\beta) \geq \varepsilon \big\} \to_P 0.$$

Now there exist $\varepsilon, \delta > 0$ such that $\inf\{\sigma(\beta)\colon |\beta| \leq \delta\} \geq \varepsilon$ and so

$$\sup_y \big\{ |\hat{Z}_n(y,\beta) - Z_n(y,\beta)| \colon |\beta| \leq \delta \big\} \to_P 0.$$

On the other hand, from (4.5)

$$\sup_y \big\{ |\hat{Z}_n(y,\beta) - Z_n(y,\beta)| \colon \delta \leq |\beta| \big\} \to_P 0$$

and so

$$\sup\big\{ |\hat{Z}_n(y,\beta) - Z(y,\beta)| \big\} \to_P 0. \tag{4.6}$$

Similarly,

$$\sup\big\{ |\hat{Z}_n^*(x,\beta) - Z^*(x,\beta)| \big\} \to_P 0, \tag{4.7}$$

where

$$\hat{Z}_n^*(x,\beta) = \sqrt{n}\left(\hat{F}_1(x) - F_1(x) - \int \Phi\left(\frac{\beta x - y + \hat{\mu}(\beta)}{\hat{\sigma}(\beta)}\right) d\big(\hat{F}_2(y) - F_2(y)\big)\right)$$

and Z^* is an appropriately defined Gaussian process. A weak consequence of (4.6) and (4.7) is that for all $\varepsilon > 0$,

$$\inf\big\{\Delta_n(\beta)\colon \varepsilon \leq |\beta^2 - \beta_0^2|\big\} \to_P \infty$$

and

$$\Delta_n(\beta_0) = O_p(1).$$

Therefore, by Proposition 4.1

$$\min\big\{|\beta_n^*(0) - \beta_0|, |\beta_n^*(0) + \beta_0|\big\} \to_P 0.$$

Since $Y - \mu(\beta) - \beta X$ is normal if and only if $\beta = \beta_0$, we conclude that $\tilde{\beta}_n$ is consistent.

We need to distinguish several cases for $n^{1/2}$-consistency:

(a) $|\beta_0| \geq \frac{3}{2}\delta_0$, $\sigma^2(\beta_0) > 0$;
(b) $|\beta_0| \geq \frac{3}{2}\delta_0$, $\sigma^2(\beta_0) = 0$;
(c) $\frac{1}{2}\delta_0 \leq |\beta_0| < \frac{3}{2}\delta_0$;
(d) $|\beta_0| \leq \frac{1}{2}\delta_0$.

(a) Suppose also that $\mathrm{Var}(Y) > \beta_0^2 \mathrm{Var}(X)$. Then, by (4.4) and (4.5)

$$\Delta_n(\beta) = \int \left| \sqrt{n} \left(F_2(y) - \int \Phi\left(\frac{y - \beta x - \mu(\beta)}{\sigma(\beta)} \right) dF_1(x) \right) + Z(y, \beta) \right|^2 \phi(y)\, dy$$
$$+ Q_n(\beta),$$

where

$$\sup\{ |Q_n(\beta)| : |\beta - \beta_0| \le \varepsilon_n \} = o_p(1).$$

Now, under these conditions,

$$\frac{\partial}{\partial \beta} \int \Phi\left(\frac{y - \beta x - \mu(\beta)}{\sigma(\beta)} \right) dF_1(x)$$
$$= -\sigma^{-1}(\beta) \int \phi\left(\frac{y - \beta x - \mu(\beta)}{\sigma(\beta)} \right) \Big(x - E(X) - \beta \sigma^{-2}(\beta)$$
$$\times (y - \beta x - \mu(\beta)) \mathrm{Var}\, X \Big)\, dF_1(x),$$

$$\frac{\partial}{\partial \beta} \int \Phi\left(\frac{y - \beta x - \mu(\beta)}{\sigma(\beta)} \right) dF_1(x) \Big|_{\beta_0} \tag{4.8}$$
$$= -\beta_0^{-1}((y - EY) f_2(y) + \mathrm{Var}\, Y f_2'(y)),$$

which cannot vanish identically as a function of y unless Y is normal (i.e., $\beta_0 = 0$ or G_0 is normal). Moreover, the derivative in (4.8) is bounded as a function of y and continuous in β. We can conclude that $\tilde{\beta}_n$ is $n^{1/2}$-consistent in this case. This follows since $\Delta_n(\beta_0) = O_p(1)$ and

$$\Delta_n(\tilde{\beta}_n) \ge \int \left(Z(y, \beta_0) + n^{1/2}(\beta_n - \beta_0) c(y) \right)^2 \phi(y)\, dy + o_p(1),$$

where $c(y)$ is the derivative in (4.6). Unboundedness of $n^{1/2}(\tilde{\beta}_n - \beta_0)$ leads to a contradiction since $c(y)$ does not vanish identically.

Case (a) with $\mathrm{Var}(Y) < \beta_0^2 \mathrm{Var}(X)$ is dealt with similarly using Z^*.

(b) If $\sigma(\beta_0) = 0$, calculate (taking $\beta_0 > 0$)

$$\lim_{\beta \to \beta_0} (\beta - \beta_0)^{-1} \left[\int \phi\left(\frac{y - \beta x - \mu(\beta)}{\sigma(\beta)} \right) dF_1(x) - F_1\left(\frac{y - \mu(\beta_0)}{\beta_0} \right) \right]$$
$$= \lim_{\beta \to \beta_0} (\beta - \beta_0)^{-1} \int \Big(F_1((y - \mu(\beta)$$
$$- z\sigma(\beta))/\beta) - F_1((y - \mu(\beta_0))/\beta_0) \Big)\, d\Phi(z)$$
$$= -\frac{y - EY}{\beta_0^2} f_1((y - \mu(\beta_0))/\beta_0) - \mathrm{Var}\, X f_1'((y - \mu(\beta_0))/\beta_0).$$

Again this expression cannot vanish identically in y unless F_1 and hence G_0 is normal. Boundedness in y and continuity in β again hold. (i) and case (b) follow.

(c) In this range since $\tilde{\beta}_n$ is consistent, we are driven to minimizing either $\Delta_n(\beta, 0)$ or $\Delta_n(\beta, 2\delta)$. In the first case, we are minimizing at $|\beta_0| > \delta/2$ and get $n^{1/2}$-consistency. In the second case, after reparametrization, we again minimize at $\delta/2 \le \beta_0 \le 7\delta/2$ and again get $n^{1/2}$-consistency.

(d) In this range since $\tilde{\beta}_n$ is consistent, we minimize $\Delta_n(\beta, 2\delta)$ with probability tending to 1. But after reparametrizing this corresponds to minimizing at $\beta_0 \ge 3\delta/2$ and we again get $n^{1/2}$-consistency. □

NOTES.

(1) For cases (ii) and (iii) of Theorem 2.2 we need to check that convergence in our arguments holds uniformly for sequences with $\|P_n - P_0\| \to 0$, $\int x^2\, dG_n(x) \to \int x^2\, dG_0(x)$, where $\|\cdot\|$ is total variation. A careful examination of the argument shows that for consistency, we need only check that

$$\bar{Y} \to_{P_n} E_0(Y), \qquad \hat{\sigma}_x^2 \to_{P_n} \mathrm{Var}_{P_0}(X),$$
$$\bar{X} \to_{P_n} E_0(X), \qquad \hat{\sigma}_y^2 \to_{P_n} \mathrm{Var}_{P_0}(Y).$$

For $n^{1/2}$-consistency, the derivatives in (4.8) and (4.9) are now evaluated at $\beta_{0n} \leftrightarrow P_n$ and depend on the marginals of T, $F_{1n} \leftrightarrow P_n$ with $\|F_{1n} - F_{10}\| \to 0$ and $F_{10} \leftrightarrow P_0$ non-Gaussian. The derivatives still converge to that for F_{10} uniformly for β bounded and are bounded uniformly in y, since $\sup_n \int |x|\, dF_{1n} < \infty$. The argument can now be made at the limit F_{10} as before.

(2) Under the restricted Gaussian error model the same argument yields that $\tilde{\beta}_{na}$ is $n^{1/2}$-consistent.

We now proceed to study the correction term which gives efficiency.

PROPOSITION 4.3. *Whatever be G_0*

$$\left|\frac{\omega_0'}{\omega_0}(t)\right| \le \tilde{\sigma}^{-2}(\theta_0)\left(|t| + \int |\eta| G_0(d\eta)\right). \tag{4.9}$$

PROOF. By a standard Laplace transform theorem, writing $\tilde{\sigma}$ for $\tilde{\sigma}(\theta_0)$,

$$\frac{\omega_0'}{\omega_0}(t) = \tilde{\sigma}^{-2}\, \frac{\int (\eta - t)\phi(\tilde{\sigma}^{-1}(t-\eta))G_0(d\eta)}{\int \phi(\tilde{\sigma}^{-1}(t-\eta))G_0(d\eta)},$$

$$\begin{aligned}\left|\int (\eta - t)\phi((\tilde{\sigma}^{-1}(t-\eta)))G_0(d\eta)\right| &\le \int |\eta - t|\phi((\tilde{\sigma}^{-1}(t-\eta)))G_0(d\eta)\\ &\le \int |\eta - t| G_0(d\eta) \int \phi((\tilde{\sigma}^{-1}(t-\eta)))G_0(d\eta),\end{aligned}$$

by an inequality of Chebyshev [Hardy, Littlewood and Pólya (1952), page 43] since $\phi(t)$ is decreasing for $t \ge 0$. □

PROPOSITION 4.4. *Suppose $H_n \to H$ weakly and $\int x^2\, dH_n(x) \to \int x^2\, dH(x)$. Then*

$$I(H_n * \Phi) \to I(H * \Phi),$$

where I denotes Fisher information for location.

PROOF. By dominated convergence for all t

$$H_n * \phi(t) \to H * \phi(t),$$
$$[H_n * \phi]'(t) \to [H * \phi]'(t).$$

By Proposition 4.3

$$\frac{|[H_n * \phi]'(t)|^2}{[H_n * \phi]} \le V(t, H_n), \tag{4.10}$$

where

$$V(t, H) = 4[H * \phi](t)\left(t^2 + \int \eta^2 H(d\eta)\right).$$

But

$$V(t, H_n) \to V(t, H) \quad \text{for all } t$$

and

$$\int V(t, H_n)\, dt = 8\int \eta^2 H_n(d\eta) + 4 \to 8\int \eta^2 H(d\eta) + 4 = \int V(t, H)\, dt.$$

The sequence in (4.10) is uniformly integrable and the result follows. □

PROPOSITION 4.5. *Let*

$$\omega_{0n}(t) = \int \omega_0(t - \sigma_n s)\lambda(s)\, ds + c_n. \tag{4.11}$$

Then if we write T_i for $T_i(\theta_0)$,

$$E\left(\frac{\hat{\omega}_n'}{\hat{\omega}_n}(T_1) - \frac{\omega_{0n}'}{\omega_{0n}}(T_1)\right)^2 \to 0, \tag{4.12}$$

$$E\left(\frac{\omega_{0n}'}{\omega_{0n}}(T_1) - \frac{\omega_0'}{\omega_0}(T_1)\right)^2 \to 0. \tag{4.13}$$

PROOF. We repeatedly use the inequalities

$$|\omega_{0n}^{(i)}| \le \sigma_n^{-i}\omega_{0n}, \qquad \omega_{0n} \le \sigma_{0n}^{-1}.$$

Write

$$\frac{\hat{\omega}_n'}{\hat{\omega}_n}(T_1)\frac{\omega_{0n}'}{\omega_{0n}}(T_1) = \frac{n^{-1}\sum_{j=1}^n\left[\lambda_n'(T_1 - T_j) - \omega_{0n}'(T_1)\right]}{\hat{\omega}_n}$$
$$- \frac{\omega_{0n}'(T_1)}{\omega_{0n}\hat{\omega}_n}\left(\frac{1}{n}\sum_{j=1}^n \lambda_n(T_1 - T_j) - \omega_{0n}(T_1)\right).$$

The first term has L_2 norm bounded by

$$c_n n^{-1/2}E^{1/2}\left([\lambda_n']^2(T_1 - T_2)\right) = O\left(c_n^{-1}\sigma_n^{-2}n^{-1/2}\right).$$

The second term is similarly norm bounded by

$$O\left(c_n^{-1}\sigma_n^{-2}n^{-1/2}\right)$$

and (4.12) follows.

For (4.13) note that, for all t, by dominated convergence,

$$\frac{\omega_{0n}'}{\omega_{0n}}(t) \to \frac{\omega_0'}{\omega_0}(t). \tag{4.14}$$

Without loss of generality, take $\tilde{\sigma}(\theta_0) = 1$. Then

$$\omega_{0n}(t) = \int \phi(t-\eta)\, d(G_0 * \lambda_n)(\eta) + c_n,$$

$$\omega_{0n}'(t) = \int \omega_0'(t - \sigma_n s)\lambda(s)\, ds.$$

By Proposition 4.3 we get

$$\frac{[\omega_{0n}']^2}{\omega_{0n}^2}(t) \le 2\left(t^2 + \int \eta^2\, dG_{s0} * \lambda_n(\eta)\right).$$

But

$$\int t^2 \omega_0(t)\, dt < \infty,$$

so that by dominated convergence and (4.14)

$$\int \left(\frac{\omega_{0n}'}{\omega_{0n}}\right)^2 (t)\omega_0(t)\, dt \to \int \frac{[\omega_0']^2}{\omega_0}(t)\, dt.$$

L_2 convergence of ω_{0n}'/ω_{0n} to ω_0'/ω_0 follows. □

Proposition 4.6. *For sequences $\{P_n\}, \{c_n\}, \{\nu_n\}$ as in Theorem* 2.2(ii), *and all M finite,*

$$\sup\left\{\left|n^{-1/2}\sum_{i=1}^n U_i(\theta)\left(\frac{\hat{\omega}_n'}{\hat{\omega}_n}(T_i(\theta),\theta) - \frac{\omega_{0n}'}{\omega_{0n}}(T_i(\theta))\right)\right| : n^{1/2}|\theta - \theta_{0n}| \le M\right\} \to_{P_n} 0, \tag{4.15}$$

where $\theta_{0n} \leftrightarrow P_{0n}$,

$$\begin{aligned}\sup\Bigg\{n^{1/2}\Bigg|\frac{1}{n}\sum_{i=1}^n \Bigg\{U_i(\theta)\left(T_i(\theta) - E_{P_n}(T_i(\theta)) + I_{0n}^{-1}\frac{\omega_{0n}'}{\omega_{0n}}(T_i(\theta))\right) \\ - U_i(\theta_{0n})\left(T_i(\theta_{0n}) - E_{P_n}(T_i(\theta_{0n})) - I_{0n}^{-1}\frac{\omega_{0n}'}{\omega_{0n}}(T_i(\theta_{0n}))\right)\Bigg\} \\ + I_{bn}(\theta - \beta_{0n})\Bigg| : n^{1/2}|\theta - \theta_{0n}| \le M\Bigg\} \to_{P_n} 0.\end{aligned} \tag{4.16}$$

This proposition reduces the proof of case (ii) to establishing that if $U_i \triangleq U_i(\theta_{0n})$, $T_i \triangleq T_i(\theta_{0n})$

$$(4.17)\quad \mathbf{L}_{P_0}\left(n^{-1/2}\sum_{i=1}^{n}\left\{U_i(T_i - E_{P_0}(T_i)) + I_{0n}^{-1}\frac{\omega'_{0n}}{\omega_{0n}}(T_i)\right\}\right) \to \mathbf{N}(0, I_b^{-1}(P_0))$$

and

$$(4.18)\qquad n^{-1}\sum_{i=1}^{n}\left(\frac{\omega'_{0n}}{\omega_{0n}}\right)^2(T_i) \to_{P_n} I_0(P_0),$$

$$(4.19)\qquad n^{-1}\sum_{i=1}^{n}U_i^2\left(T_i + I_0^{-1}(P_0)\frac{\omega'_{0n}}{\omega_{0n}}(T_i)\right)^2 \to_{P_n} I_b(P_0).$$

All three claims follow since

$$\mathbf{L}_{P_n}(U_1, T_1) \to \mathbf{L}_{P_0}(U_1, T_1),$$
$$\frac{\omega'_{0n}}{\omega_{0n}}(t) \to \frac{\omega'_0}{\omega_0}(t), \quad \text{for all } t,$$

and $E_{P_n}(U_1^2)$, $E_{P_n}(T_1^2)$, $\int([\omega'_{0n}]^2/\omega_{0n})(t)\,dt$ all converge to the appropriate limits under P_0. The last claim is a consequence of Proposition 4.4.

Proof of Proposition 4.6. Denote the (random) functions in absolute values in (4.15) by

$$Q_n(\Delta), \quad \text{where } \Delta = (\theta - \theta_{0n})n^{1/2}.$$

Now

$$(4.20)\qquad Q_n(0) \to_{P_n} 0$$

by Proposition 4.5.

Write

$$Q_{1n}(\Delta) = n^{-1}\sum_{i=1}^{n}T_i\left(\frac{\hat\omega'_n}{\hat\omega_n}(T_i(\theta), \theta) - \frac{\omega'_{0n}}{\omega_{0n}}(T_i(\theta))\right),$$

$$Q_{2n}(\Delta) = n^{-1/2}\sum_{i=1}^{n}U_{in}\left(\frac{\hat\omega'_n}{\hat\omega_n}(T_i(\theta), \theta) - \frac{\omega'_{0n}}{\omega_{0n}}(T_i(\theta))\right).$$

It is easy to see that for (4.15) we need only check that

$$(4.21)\qquad \sup\{|Q_{in}(\Delta)|: |\Delta| \le M\} \to_{P_n} 0, \qquad i = 1,2.$$

Throughout this calculation we write $\lambda_n = \lambda_{\nu_n}$ and repeatedly use

$$\hat\omega_n \ge c_n, \qquad \omega_{0n} \ge c_n, \qquad |\lambda_n^{(i)}| \le \nu_n^{-i}\lambda_n.$$

We begin with $i = 1$. Let

$$V_{1n}(\Delta) = \frac{\hat\omega'_n}{\hat\omega_n}(T_1(\theta), \theta) - \frac{\omega'_{0n}}{\omega_{0n}}(T_1(\theta), \theta).$$

By Cauchy-Schwarz and uniform integrability of T_1^2 (as P_n varies), it is enough to check that

$$E \sup_{\Delta} (V_{1n}(\Delta))^2 = O\big(n^{-1}\nu_n^{-4}(c_n^{-2} + \log n)\big). \tag{4.22}$$

Note first that

$$|V_{1n}(0)| \le c_n^{-1}|\hat{\omega}_n'(T_1, \theta_{0n}) - \omega_{0n}'(T_1)| + c_n^{-1}\nu_n^{-1}|\hat{\omega}_n(T_1, \theta_{0n}) - \omega_{0n}(T_1)|. \tag{4.23}$$

Let $\hat{F}_n$ be the empirical distribution function of $T_1, \ldots, T_n$ and F its expectation. Then

$$\begin{aligned} |\hat{\omega}_n'(t, \theta_{0n}) - \omega_{0n}'(t)| &= O\big(n^{-1}\sigma_n^{-2}\big) + n^{-1}\sum_{i=2}^{n}[\lambda_n'(t - T_i) - E\lambda_n'(t - T_i)] \\ &= O\big(n^{-1}\sigma_n^{-2}\big) + O(1)\int\big(\hat{F}_n(s) - F(s)\big)\lambda_n''(t - s)\,ds \\ &\le O\big(n^{-1}\sigma_n^{-2}\big) + O(1)\sup_s|\hat{F}_n(s) - F(s)|\int|\lambda''(s)|\,ds, \end{aligned}$$

where the 0 terms are nonstochastic and independent of t. A similar bound holds for the second term in (4.23) and hence

$$EV_{1n}^2(0) = O\big(n^{-1}\nu_n^{-4}c_n^{-2}\big). \tag{4.24}$$

Next we write

$$T_i(\theta) = T_i + \frac{a}{\sqrt{n}}U_i + \frac{b}{\sqrt{n}}T_i,$$

so that a, b are well defined functions of Δ and note that

$$\begin{aligned} \frac{\partial}{\partial a}V_{1n}(\Delta) = n^{-1/2}\Bigg\{&\frac{\Sigma(U_1 - U_j)\lambda_n''(T_1(\theta) - T_j(\theta))}{nc_n + \Sigma\lambda_n(T_1(\theta) - T_j(\theta))} - \frac{\hat{\omega}_n'}{\hat{\omega}_n}(T_1(\theta), \theta) \\ &\times\frac{\Sigma(U_1 - U_j)\lambda_n'(T_1(\theta) - T_j(\theta))}{nc_n + \Sigma\lambda_n(T_1(\theta) - T_j(\theta))} - U_1\left(\frac{\omega_{0n}'}{\omega_{0n}}\right)'(T_1(\theta))\Bigg\}. \end{aligned}$$

Therefore,

$$\begin{aligned} &E\sup\left\{\left|\frac{\partial}{\partial a}V_{1n}(\Delta)\right| : (\Delta) \le M\right\}^2 \\ &\quad\le C(M)n^{-1}\nu_n^{-4}E\Big(\max_j(U_1 - U_j)^2 + U_1^2\Big) = O\left(\frac{\log n}{n}\sigma_n^{-4}\right). \end{aligned} \tag{4.25}$$

Similarly, we can bound

$$\begin{aligned} \left|\frac{\partial}{\partial b}V_{1n}(\Delta)\right| \le n^{-1/2}\Bigg|&\frac{\Sigma(T_1 - T_j)\lambda_n''(T_1(\theta) - T_j(\theta))}{\hat{\omega}_n(T_1(\theta), \theta)} \\ &-\frac{\hat{\omega}_n'}{\hat{\omega}_n^2}(T_1(\theta), \theta)\sum|T_1 - T_j|\lambda_n'(T_1(\theta) - T_j(\theta)) \\ &-T_1\left(\frac{\omega_{0n}'}{\omega_{0n}}\right)'(T_1(\theta))\Bigg|. \end{aligned} \tag{4.26}$$

Representing $T_i = kT_i(\theta) + (c/\sqrt{n})U_i$, $k \to 1$, we can bound (4.26) by

$$An^{-1/2}\left\{ \nu_n^{-2} \frac{\Sigma|T_1(\theta) - T_j(\theta)|\lambda_n(T_1(\theta) - T_j(\theta))}{nc_n + \Sigma\lambda_n(T_1(\theta) - T_j(\theta))} + n^{-1/2}\left|\frac{\partial V_{1n}}{\partial a}(\Delta)\right| + \nu_n^{-2}(|U_1| + |T_1|)\right\}.$$

Representing $T_i = kT_i(\theta) + (c/\sqrt{n})U_i$, $k \to 1$, we can bound (4.26) by

$$An^{-1/2}\left\{ \nu_n^{-2} \frac{\Sigma|T_1(\theta) - T_j(\theta)|\lambda_n(T_1(\theta) - T_j(\theta))}{nc_n + \Sigma\lambda_n(T_1(\theta) - T_j(\theta))} + n^{-1/2}\left|\frac{\partial V_{1n}}{\partial a}(\Delta)\right| + \nu_n^{-2}(|U_1| + |T_1|)\right\},$$

for a constant A depending on M only. Since $\lambda_\nu(|t|)$ is decreasing, the first term in curly brackets is bounded using the Chebyshev inequality by

$$n^{-1}\sum|T_1(\theta) - T_j(\theta)|. \tag{4.27}$$

Since (4.27) is bounded by

$$B\left\{n^{-1}\sum\left(|T_j| + |T_1| + n^{-1/2}(|U_j| + |U_1|)\right)\right\},$$

for B depending on M only, we obtain

$$E\sup\left\{\left|\frac{\partial V_{1n}}{\partial b}(\Delta)\right|^2 : |\Delta| \le M\right\} = O(n^{-1}\nu_n^{-4}\log n). \tag{4.28}$$

Combining (4.24), (4.25) and (4.28), we get (4.21) for $i = 1$.

The proof of (4.21) for $i = 2$ is similar, but more complicated using the almost independence of $U_i(\theta)$, $T_i(\theta)$.

First, since $\hat{\omega}_n(\cdot, \theta_0)$ does not depend on the U_i,

$$\begin{aligned} EQ_{2n}^2(0) &= EU_1^2 E(V_{1n}^2(0)) \\ &= O(n^{-2}\nu_n^{-4}c_n^{-2}). \end{aligned} \tag{4.29}$$

Next,

$$\begin{aligned} \frac{\partial Q_{2n}}{\partial a}(\Delta) &= \frac{1}{n}\sum_j U_j^2\left(\left(\frac{\hat{\omega}_n'}{\hat{\omega}_n}\right)'(T_j(\theta), \theta_0) - \left(\frac{\omega_{0n}'}{\omega_{0n}}\right)'(T_j(\theta))\right) \\ &\quad + n^{-1}\sum_i U_i \frac{\Sigma_j U_j \lambda_n''(T_i(\theta) - T_j(\theta))}{nc_n + \Sigma\lambda_n(T_i(\theta) - T_j(\theta))} \\ &\quad - n^{-1}\sum_i U_i \frac{\hat{\omega}_n'}{\hat{\omega}_n}(T_i(\theta), \theta)\frac{\Sigma_j U_j \lambda_n'(T_i(\theta) - T_j(\theta))}{nc_n + \Sigma\lambda_n(T_i(\theta) - T_j(\theta))} \\ &= R_{1n}(\Delta) + R_{2n}(\Delta) + R_{3n}(\Delta), \quad \text{say}. \end{aligned}$$

By arguing as for (4.22)

$$\sup\{ER_{1n}^2(\Delta)\colon |\Delta| \le M\} = O\big(n^{-1}\nu_n^{-6}\big(c_n^{-2} + \log n\big)\big).$$

The additional ν_n^{-2} comes from the third derivatives in λ_n we have to deal with.

To deal with R_{2n} and R_{3n}, note that we can define $c(\theta)$ such that the Gaussian random variable

$$(4.30)\qquad \tilde{U}_i(\theta) = U_i + \frac{c(\theta)}{\sqrt{n}}(T_i - X_i')$$

is independent of $T_i(\theta)$. This follows since $T_i(\theta)$ is a linear combination of X_i' and the Gaussian variables U_i and $T_i - X_i'$, both of which are independent of X_i'. Using (4.30)

$$\begin{aligned} ER_{2n}^2(\Delta) &\le 4E\left(n^{-2}\sum_{i,j}\tilde{U}_i\tilde{U}_j(\theta)\frac{\lambda_n''(T_i(\theta) - T_j(\theta))}{\hat{\omega}_n(T_i(\theta),\theta)}\right)^2 \\ &\quad + 4E\left(n^{-2}\sum_{i,j}\big(U_iU_j - \tilde{U}_i\tilde{U}_j(\theta)\big)\frac{\lambda_n''(T_i(\theta) - T_j(\theta))}{\hat{\omega}_n(T_i(\theta),\theta)}\right)^2 \\ &= O\big(n^{-1}\nu_n^{-4}\big) + O\big(n^{-1}\log n\nu_n^{-4}\big), \end{aligned}$$

since

$$E\tilde{U}_i^2(\theta) = O(1),$$

$$E\max\big(\tilde{U}_i\tilde{U}_j(\theta) - U_iU_j\big)^2 = O\big(n^{-1}\log n\big).$$

We can bound $ER_{3n}^2(\Delta)$ similarly to get

$$(4.31)\qquad \sup\left\{E\left(\frac{\partial}{\partial a}Q_{2n}(\Delta)\right)^2 : |\Delta| \le M\right\} = O\big(n^{-1}\nu_n^{-6}\big(c_n^{-2} + \log n\big)\big).$$

Finally, we need to study $(\partial/\partial b)Q_{2n}(\Delta)$. It is possible to pass from the bound on $E((\partial/\partial a)Q_{2n}(\Delta))^2$ to the bound on $E((\partial/\partial b)Q_{2n}(\Delta))^2$ as was done in the passing from the bound on $(\partial/\partial a)V_{1n}(\Delta)$ to the bound on $(\partial/\partial b)V_{1n}(\Delta)$. We conclude

$$(4.32)\qquad \sup\left\{E\left(\frac{\partial}{\partial b}Q_{2n}(\Delta)\right)^2 : |\Delta| \le M\right\} = O\big(n^{-1}\sigma_n^{-6}\big(c_n^{-2} + \log n\big)\big).$$

If we combine (4.31) and (4.32) with (4.29), we get by the standard Billingsley–Chentsov fluctuation inequalities [Billingsley (1968)],

$$\sup\{|V_{2n}(\Delta)|\colon |\Delta| \le M\} = O_{P_n}\big(n^{-1}\sigma_n^{-6}\big(c_n^{-2} + \log n\big)\big).$$

The proof of (4.15) is complete.

We now prove (4.16). Let

$$W_n(\Delta) = n^{-1/2}\bar{\sigma}(\theta)\sum_{i=1}^{n}U_i(\theta)\left(T_i(\theta) - E_{P_n}(T_i(\theta)) + I_{0n}^{-1}\frac{\omega_{0n}'}{\omega_{0n}}(T_i(\theta))\right),$$

where $\theta = \theta_0 + \Delta n^{-1/2}$, $\Delta = (\Delta_1, \dots, \Delta_4)$, $\Delta_1 = \beta$, etc. Claim (4.16) is equivalent to

$$\sup\left\{\left|W_n(\Delta) - W_n(0) - \sum_{j=1}^{4} \frac{\partial W_n(0)}{\partial \Delta_j}\Delta_j\right| : |\Delta| \le M\right\} \to_{P_n} 0 \tag{4.33}$$

and

$$\left|\frac{\partial W_n(0)}{\partial \Delta_j} - I_{bn}\bar{\sigma}(\theta_0)\delta_{1j}\right| \to_{P_n} 0, \qquad j = 1, \dots, 4. \tag{4.34}$$

Now,

$$\frac{\partial W_n(0)}{\partial \Delta_1} = n^{-1} \sum_{i=1}^{n} \left[X_i\left(T_i + I_{0n}^{-1}\frac{\omega_{0n}'}{\omega_{0n}}(T_i)\right) + U_i(\gamma_1 U_i + \gamma_2(T_i - ET_i)) \times\left(1 + I_{0n}^{-1}\left(\frac{\omega_{0n}'}{\omega_{0n}}\right)'(T_i)\right)\right],$$

for suitable γ_1, γ_2, the laws of the summands converge to $\mathbf{L}_0(A)$, where

$$A = X\left(T + I_0^{-1}\frac{\omega_0'}{\omega_0}(T)\right) + U(\gamma_1 U + \gamma_2(T - E_0 T))\left(1 + I_0^{-1}\left(\frac{\omega_0'}{\omega_0}\right)'(T)\right),$$

and the summands are uniformly integrable (P_n) by Proposition 4.4. Therefore,

$$\frac{\partial W_n}{\partial \Delta_1}(0) \to_{P_n} E_0(A) = I_b\bar{\sigma}(\theta_0),$$

after some computation. A similar argument establishes (4.34) for $j > 1$. For (4.33) we check that for $1 \le j \le k \le 4$,

$$\sup\left\{\left|\frac{\partial^2 W_n}{\partial \Delta_j\, \partial \Delta_k}(\Delta)\right| : |\Delta| \le M\right\} \to 0. \tag{4.35}$$

We give the argument for a typical term, $\Delta_3 \leftrightarrow \nu_1$,

$$\frac{\partial^2 W_n}{\partial \Delta_3^2} = n^{-3/2} \sum_{i=1}^{n} \sigma(\theta) U_i(\theta) I_{0n}^{-1} X_{in}^2 \left(\frac{\omega_{0n}'}{\omega_{0n}}\right)''(T_i(\theta)). \tag{4.36}$$

Since $|\omega_{0n}^{(i)}/\omega_{0n}| \le \sigma_n^{-i}$, we bound (4.35) uniformly in $|\Delta| \le M$ by

$$n^{-1/2}\sigma_n^{-2}O(1)\left\{n^{-1} \sum_{i-1}^{n} |U_i|(T_i^2 + U_i^2) + n^{-3/2} \sum_{i=1}^{n} |T_i|^3\right\}. \tag{4.37}$$

Since T_i^2 are uniformly integrable under P_n,

$$n^{-1/2} \max_i |T_i| \to_{P_n} 0. \tag{4.38}$$

Claim (4.35) for $j = k = 3$ follows from (4.37) and (4.38). The other terms are dealt with similarly and the result follows.

Proposition 4.6 establishes claim (ii) of the theorem. For part (iii) note that Proposition 4.6 shows that if β_n^* is $n^{1/2}$-consistent so is $\hat{\beta}_n(\beta_n^*)$ and, in fact,

$$\hat{\beta}_n(\beta_n^*) = \beta_{0n} + n^{-1} \sum_{i=1}^{n} \tilde{I}_b(X_i, P_n) + o_{P_n}(n^{-1/2}).$$

Therefore, taking β_n^* successively as $\hat{\beta}_{0n}, \hat{\beta}_{1n}, \dots,$ we get

$$\hat{\beta}_{in} - \hat{\beta}_{1n} = o_{P_n}(n^{-1/2})$$

and claim (iii) follows. Claim (iv) is established in exactly the same way as claims (i)–(iii). □

PROPOSITION 4.7. *The efficiency of $\hat{\beta}_P$ under model (Identity, Φ), I_c/I_a, satisfies*

$$I_c/I_a \geq \left(1 + \sigma^2/(\beta^2 + 1)\left(\operatorname{Var}(X') + \sigma^2\right)\right)^{-1}.$$

PROOF.

$$\begin{aligned} I_a/I_c &= [\operatorname{Var}(X')]^{-2}\operatorname{Var}(T)\left[\operatorname{Var}(T) - 2\tilde{\sigma}^2 + \tilde{\sigma}^4 I_0\right] \\ &= 1 + \tilde{\sigma}^4\left(I_0\operatorname{Var}(T) - 1\right)/\left(\operatorname{Var}(X')\right)^2, \end{aligned}$$

since $\operatorname{Var}(T) = \operatorname{Var}(X') + \tilde{\sigma}^2$. Since T is, in general, an inefficient estimate of η in the location model $T = \eta + \varepsilon$ we must have $\tilde{\sigma}^2 \geq I_0^{-1}$ so that

$$\begin{aligned} I_a/I_c - 1 &\leq \tilde{\sigma}^4\left(\operatorname{Var}(T)/\tilde{\sigma}^2 - 1\right)/\left(\operatorname{Var}(X')\right)^2 \\ &= \tilde{\sigma}^2/\operatorname{Var}(X') = \sigma^2/(\beta^2 + 1)\operatorname{Var}(X') \end{aligned}$$

and the result follows. □

Acknowledgment. We thank Cliff Spiegelman for helpful discussions.

REFERENCES

ANDERSON, T. W. (1984). Estimating linear statistical relationships. *Ann. Statist.* **12** 1–45.

BEGUN, J. M., HALL, W. J., HUANG, W.-M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432–452.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

CSÖRGŐ, M. (1981). *Strong Approximations in Probability and Statistics.* Academic, New York.

EFRON, B. (1977). The efficiency of Cox's likelihood function for censored data. *J. Amer. Statist. Assoc.* **72** 557–565.

GLESER, L. (1981). Estimation in a multivariate "errors in variables" regression model: Large sample results. *Ann. Statist.* **9** 24–44.

HARDY, G. H., LITTLEWOOD, J. E. and PÓLYA, G. (1952). *Inequalities*, 2nd ed. Cambridge Univ. Press.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory.* Springer, New York.

KENDALL, M. G. and STUART, A. (1979). *The Advanced Theory of Statistics* **2**, 4th ed. Hafner, New York.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906.

KOSHEVNIK, YU. A. and LEVIT, B. YA. (1976). On a non-parametric analogue of the information matrix. *Theory Probab. Appl.* **21** 738–753.

NEYMAN, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In *Probability and Statistics* (U. Grenander, ed.) 213–234. Wiley, New York.

NEYMAN, J. and SCOTT, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32.

PFANZAGL, J. (1982). *Contributions to a General Asymptotic Statistical Theory. Lectures Notes in Statist.* **13**. Springer, New York.

REIERSØL, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica* **18** 375–389.

RUBIN, H. (1956). Uniform convergence of random functions with applications to statistics. *Ann. Math. Statist.* **27** 200–203.

SPIEGELMAN, C. (1979). On estimating the slope of a straight line when both variables are subject to error. *Ann. Statist.* **7** 201–206.

STEIN, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 187–195. Univ. of California Press.

COURANT INSTITUTE OF MATHEMATICAL SCIENCES
NEW YORK UNIVERSITY
251 MERCER STREET
NEW YORK, NEW YORK 10012

DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY
JERUSALEM 91905
ISRAEL

The Annals of Statistics
1990, Vol. 18, No. 2, 925–938

ACHIEVING INFORMATION BOUNDS IN NON AND SEMIPARAMETRIC MODELS[1]

BY Y. RITOV AND P. J. BICKEL

The Hebrew University of Jerusalem and University of California, Berkeley

We consider in this paper two widely studied examples of nonparametric and semiparametric models in which the standard information bounds are totally misleading. In fact, no estimators converge at the $n^{-\alpha}$ rate for any $\alpha > 0$, although the information is strictly positive "promising" that $n^{-1/2}$ is achievable. The examples are the estimation of $\int p^2$ and the slope in the model of Engle et al. A class of models in which the parameter of interest can be estimated efficiently is discussed.

1. Introduction. Consider the standard simple random sampling model on a sample space $\mathbf{X}$: $X_1, \dots, X_n$ i.i.d. according to $P \in \mathbf{P}$, a set of probability measures on $\mathbf{X}$ dominated by μ. Let p denote the density of P and θ: $\mathbf{P} \to R$ be a parameter. Suppose $\mathbf{P}$ is a regular parametric model, that is,

1. $\mathbf{P} = \{P_{(\theta,\eta)}: \theta \in R,\ \eta \in R^m\}$, where if $s(\theta,\eta) = [dP_{(\theta,\eta)}/d\mu]^{1/2}$, the map $(\theta,\eta) \to s(\theta,\eta)$ is continuously Fréchet differentiable from R^{m+1} to $L_2(\mu)$, with derivative $\dot{s}(\theta,\eta)$ an $m+1$ vector of elements of $L_2(\mu)$.
2. The Fisher information matrix, $I(\theta,\eta) = 4[\int \dot{s}_i(\theta,\eta)\dot{s}_j(\theta,\eta)\,d\mu]_{(m+1)\times(m+1)}$ (where the $\dot{s}_i$ are the components of $\dot{s}$), is nonsingular.

Then it is known [see, for example, Hájek (1972)] that if θ is identifiable it can be estimated at rate $1/\sqrt{n}$. In fact, there exist $\hat{\theta}_n$ of "maximum likelihood" type which have the property that, if I^{11} is the first element of I^{-1}, then

$$\mathbf{L}_\theta \mathbf{X}\big(n^{1/2}(\hat{\theta} - \theta)\big) \to \mathbf{N}\big(0, I^{11}(\theta,\eta)\big)$$

uniformly on compact subsets of R^{m+1} and I^{11} is the smallest asymptotic variance achievable by uniformly converging estimates.

Levit (1978), Pfanzagl (1982) and Begun, Hall, Huang and Wellner (1983) have used an idea of Stein (1956) to extend those lower bounds to $\mathbf{P}$ nonparametric or semiparametric, provided that θ is pathwise Hellinger differentiable on $\mathbf{P}$.

In this paper we investigate the question: Under the conditions of the above authors, are the bounds necessarily sharp if we drop the restriction that $\mathbf{P}$ is a regular parametric model?

Received October 1987; revised July 1989.
[1]Research supported by ONR grant N00014-80-C-0163.
AMS 1980 *subject classifications.* 62G20, 62G05.
Key words and phrases. Rate of convergence, nonparametric estimations, functionals of a density.

We begin, in Section 2, by showing in the context of two widely studied examples, estimation of $\int p^2$, and of the regression coefficient in the model of Engle, Granger, Rice and Weiss (1986) that the answer is, in general, no. In fact, the rate $n^{-1/2}$ is not even achievable pointwise. Although the arguments are specific, they can evidently be generalized to show similar results for much broader classes of parameters. A general view of these phenomena is given in Donoho and Liu (1988).

In Section 3 we show that the information bounds are valid for a general class of semiparametric models. The class includes the regular parametric models and is rich enough to contain models having essentially any tangent space structure.

2. The bounds are not sharp. The first example we consider is

$$\mathbf{P} \equiv \{P \text{ on } [0,1]\colon P \text{ absolutely continuous with density } p \le M\},$$

where M is a finite constant and,

$$\theta(p) = \int p^2(x)\,dx.$$

Since the functional $\theta(p)$ is differentiable along every Hellinger path in $\mathbf{P}$, the regularity conditions required for validity of the information bound are satisfied. This functional appears in the asymptotic variance of the Hodges–Lehmann estimator. Similar functions (the integral of the square of the derivative of the density) appear in the theory of optimal density estimation.

It is well known [Pfanzagl (1982) and Donoho and Liu (1988)] that the information bound in this case is

$$4\operatorname{Var} p(X) = 4\int \big(p(x) - \theta(p)\big)^2 p(x)\,dx. \tag{2.1}$$

Hasminskii and Ibragimov (1979), following work of Schweder (1975), exhibit an estimate $\hat\theta_n$ such that $\sqrt{n}\,(\hat\theta_n - \theta(p))/2[\operatorname{Var} p(X)]^{1/2}$ converges in law to $\mathbf{N}(0,1)$ uniformly on $\{P$ with densities p such that $\|p\|_\infty + \|p'\|_\infty \le L\}$. Yet we can establish the following.

THEOREM 1. *For any $\varepsilon > 0$, there exists a subset $\mathbf{P}_0 \subset \mathbf{P}$ (compact in the topology induced by the variational norm and having diameter less than ε) such that for every sequence of estimators $\hat\theta_n$ and every $\alpha > 0$, there exists $P \in \mathbf{P}_0$ such that*

$$\liminf_n P\Big[\big|\hat\theta_n - \theta\big| \ge n^{-\alpha}\Big] > 0. \tag{2.2}$$

A consequence of this result is that the rate of convergence on $\mathbf{P}_0$, as defined, for example, by Stone (1980), is slower than $n^{-\alpha}$ for any $\alpha > 0$. In fact, no sequence of estimators which is $n^{-\alpha}$ consistent at each point of $\mathbf{P}_0$ exists. So the information bound is totally misleading for $\mathbf{P}$.

To see what goes wrong, we consider the behaviour of a plausible type of estimator. It is proved in Pfanzagl (1982)—see also Bickel, Klaassen, Ritov and Wellner (to which we refer in the sequel as BKRW)—that if $\hat{\theta}_{\text{eff}}$ is efficient, then

$$\hat{\theta}_{\text{eff}} = \theta(p) + 2n^{-1}\sum_{i=1}^{n}(p(X_i) - \theta(p)) + o_p(n^{-1/2}).$$

The naive approach to estimating θ efficiently is to try $\tilde{\theta} = \theta(\hat{p}_n) + 2n^{-1}\sum_{i=1}^{n}[\hat{p}_n(X_i) - \theta(\hat{p}_n)]$ for $\hat{p}_n$ an estimator of the density. For simplicity, suppose $\hat{p}_n(\cdot)$ is based on an auxiliary sample. If $\tilde{\theta} = \hat{\theta}_{\text{eff}} + o_p(n^{-1/2})$, we would expect

$$E(\tilde{\theta}|\hat{p}_n) = \int p^2(x)\,dx + O_p(n^{-1/2}).$$

But,

$$\begin{aligned} E(\tilde{\theta}|\hat{p}_n) - \int p^2(x)\,dx &= 2\int \hat{p}_n(x)p(x)\,dx - \int \hat{p}_n^2(x)\,dx - \int p^2(x)\,dx \\ &= -\int (\hat{p}_n(x) - p(x))^2\,dx. \end{aligned}$$

According to Bretagnolle and Huber (1979), to have this last term be of order $n^{-1/2}$ uniformly for $p \in \mathbf{P}$ we need a Hölder condition of order at least $\frac{1}{2}$ on p in $\mathbf{P}$, viz. $|p(x) - p(y)| \le c|x - y|^{1/2}$. A positive result when p is so restricted has been obtained by Ibragimov and Haminskii (1979). This argument cannot be translated into a proof since we have considered only estimates of a particular type in the discussion of the rate at which p can be estimated. In fact, a cleverer construction [see Bickel and Ritov (1988)] shows that a Hölder condition of order $\frac{1}{4}$ suffices. However, we hope the point is clear. The calculations leading to the information bound are local. They are irrelevant to actual performance if you can't even get to within $o_p(n^{-1/4})$ of $\theta(p)$.

We begin with a simpler construction which establishes the following.

THEOREM 2. *For any sequence of estimates $\hat{\theta}_n$ there exists a compact $\mathbf{P}_0$ for which the uniform rate of convergence is slower than a_n, for any sequence $a_n \to 0$, viz.*

$$\liminf_n \sup_{\mathbf{P}_0} P\big[|\hat{\theta}_n - \theta| \ge a_n\big] > 0. \tag{2.3}$$

Note that (2.3) *implies the existence of $\varepsilon > 0$ such that*

$$\liminf_n \sup_{p_0} P\big[|\hat{\theta}_n - \theta| \ge \varepsilon\big] > 0.$$

The main idea of the proof is a "Bayesian" construction. We exhibit a sequence of prior distributions π_n, assigning mass $\frac{1}{2}$ each to finite subsets H_{0n} of $\{P: \theta(P) = 1 + \frac{4}{3}a_n\}$ and H_{1n} of $\{P: \theta(P) = 1 + \frac{16}{3}a_n\}$, whose size $k(n)\uparrow\infty$ such that the posterior probabilities of H_{1n}, H_{0n} given $X_1, \dots, X_n$ are, with probability tending to 1, still equal to $\frac{1}{2}$. More explicitly, the members p_{jln}, $l = 1, \dots, k(n)$, of H_{jn}, $j = 0, 1$, are equally likely a priori and are chosen so that, with probability tending to 1,

$$k^{-1}(n) \sum_{l=1}^{k(n)} \prod_{i=1}^{n} p_{0ln}(X_i) = k^{-1}(n) \sum_{l=1}^{k(n)} \prod_{i=1}^{n} p_{1ln}(X_i) = \prod_{i=1}^{n} p(X_i),$$

where p is the uniform distribution on $(0,1)$ (though this is inessential). Define $\mathbf{P}_0$ to be this countable collection of P_{jlm}'s together with their limit, the uniform distribution. An immediate consequence from which (2.3) follows is that,

$$\inf_{\hat{\theta}_n} \int P\left[\left|\hat{\theta}_n - \theta\right| \geq a_n\right] \pi_n(dP) \to \tfrac{1}{2},$$

and this establishes the theorem. This construction differs from similar constructions appearing in the density estimation literature where the corresponding H_{0n}, H_{1n} are simple (consist of one point).

Proof of Theorem 2. Here is the sequence of priors, the union of whose carriers is a set having the uniform distribution on $(0,1)$ as its limit. We prescribe π_n through some auxiliary variables.

(1) Let

$$\alpha_n = \begin{cases} c_n, & \text{with probability } \tfrac{1}{2}, \\ 2c_n, & \text{with probability } \tfrac{1}{2}; \end{cases}$$

the sequence $c_n \downarrow 0$ is to be chosen later.

(2) Let $\Delta_0, \dots, \Delta_m$, $m = n^3$, be independent identically distributed random variables independent of α_n and equal to ± 1 with probability $\frac{1}{2}$.

π_n is the distribution of the random density p given by

$$p\big((i+y)(m+1)^{-1}\big) = 1 + \Delta_i \alpha_n h(y), \qquad i = 0, \dots, m, 0 \leq y \leq 1,$$

where (say)

$$h(t) = \begin{cases} t, & 0 \leq t < \tfrac{1}{2}, \\ -(1-t), & \tfrac{1}{2} \leq t \leq 1. \end{cases}$$

The support of each π_n is finite and $\int |p - 1| \leq 2c_n$ with π_n probability 1, so the union of the supports of π_n is a sequence tending to the uniform distribution. Now, if P corresponds to the random p,

$$\theta(P) = \int p^2(x)\,dx = (m+1)^{-1} \sum_{i=0}^{m} \int_0^1 \big(1 + \Delta_i \alpha_n h(y)\big)^2\,dy = 1 + \frac{\alpha_n^2}{12}.$$

This construction, since $m = n^3$, has the property that the π_n probability that at most one of the observed $X_1, \dots, X_n$ will fall into any of the intervals $[i/(m+1), (i+1)/(m+1))$ is $1 - O(n^{-1})$. But one observation in a cell gives no new information on whether $\alpha_n = c_n$ or $2c_n$ and so the posterior probability,

$$\begin{aligned}(2.4)\qquad \pi_n\left\{\theta = 1 + \frac{c_n^2}{12}\middle|X_1,\dots,X_n\right\} &= \pi_n\left\{\theta = 1 + \frac{c_n^2}{3}\middle|X_1,\dots,X_n\right\}\\ &= \frac{1}{2} + o_{\pi_n}(1).\end{aligned}$$

Let $c_n = 3a_n^{1/2}$. Then (2.4) implies that

$$\inf_{\hat{\theta}} P\left(\left|\hat{\theta}_n - \theta\right| > a_n\middle|X_1, X_2,\dots,X_n\right) \to_{\pi_n} \tfrac{1}{2},$$

or, for any $\hat{\theta}_n = \hat{\theta}_n(X_1,\dots,X_n)$,

$$\int P\left[\left|\hat{\theta}_n - \theta\right| \geq a_n\right]\pi_n(dP) \to \tfrac{1}{2}.$$

Then

$$\liminf_n \sup_{\mathbf{P}_0} P\left[\left|\hat{\theta}_n - \theta\right| > a_n\right] \geq \liminf_n \int P\left[\left|\hat{\theta}_n - \theta\right| > a_n\right]\pi_n(dP) = \tfrac{1}{2}$$

and (2.3) follows. To check (2.4), note that if at most one X_i falls in each interval, the posterior distribution of $(\alpha_n, \Delta_0,\dots,\Delta_m)$ is

$$\begin{aligned}&\pi_n(\alpha, \Delta_0,\dots,\Delta_m|X_1,\dots,X_n)\\ (2.5)\qquad &= 2^{-(m+2)}\prod_{i=0}^{m}\left\{\frac{1+\Delta_i}{2}f_\alpha^+(Y_i) + \frac{1-\Delta_i}{2}f_\alpha^-(Y_i)\right\}^{\delta_i} c(X_1,\dots,X_n)\\ &= \prod_{i=0}^{m}\{1 + \Delta_i\alpha h(Y_i)\}^{\delta_i},\end{aligned}$$

where

$$f_\alpha^{\pm}(y) = 1 \pm \alpha h(y),$$

$$\delta_i = \begin{cases} 1, & \text{if there exists } X_{j_i} \in \left[\dfrac{i}{m+1}, \dfrac{i+1}{m+1}\right), \\ 0, & \text{otherwise,}\end{cases}$$

and Y_i is the fractional part of $(m+1)X_{j_i}$. By symmetry, from (2.5),

$$\pi_n(\alpha_n = c_n|X_1,\dots,X_n) = \tfrac{1}{2}$$

and (2.4) follows. □

Theorem 1 again uses a Bayesian construction. For the conclusion we cannot reduce our problem from estimation to testing but have to construct a prior distribution with infinite support whose Bayes risk for the loss function $l_n(\theta, \hat{\theta}) = 1(|\hat{\theta} - \theta| \geq a_n)$ is bounded away from 0.

PROOF OF THEOREM 1. We exhibit a $\mathbf{P}_0$ contained in the ε ball around $\mathbf{U}(0,1)$ and π_0 concentrating on $\mathbf{P}_0$ such that for all $\alpha > 0$,

$$\liminf_n \inf_{\hat{\theta}_n} \int P\left[\left|\hat{\theta}_n - \theta\right| \geq n^{-\alpha}\right]\pi_0(dP) \geq \tfrac{1}{4}. \tag{2.6}$$

Then (2.2) follows. Otherwise, we could exhibit $\alpha > 0$, $\hat{\theta}_n$ such that for all P,

$$P\left[\left|\hat{\theta}_n - \theta\right| \geq n^{-\alpha}\right] \to 0,$$

which by dominated convergence would imply

$$\int P\left[\left|\hat{\theta}_n - \theta\right| \geq n^{-\alpha}\right]\pi_0(dP) \to 0,$$

contradicting (2.6). Here is π_0. Let $\alpha_k, \Delta_k(0), \dots, \Delta_k(2^k - 1)$, $k = 1, 2, \dots$ be independent, $\alpha_k = 0$ or 1 with probability $\frac{1}{2}$, each $\Delta_k(i) = \pm 1$ with probability $\frac{1}{2}$ each. Define the random functions

$$h_k(x) = \begin{cases} \Delta_k(i), & i2^{-k} \leq x < \left(i + \frac{1}{2}\right)2^{-k}, \\ -\Delta_k(i), & \left(i + \frac{1}{2}\right)2^{-k} \leq x < (i+1)2^{-k}. \end{cases} \tag{2.7}$$

Finally, the random density p is given by

$$p(x) = 1 + \sum_{k=1}^{\infty} c_k \alpha_k h_k(x),$$

where the c_k are positive $\sum_{k=1}^{\infty} c_k < \varepsilon/2$. Note that since $\int h_i(x)\,dx = 0$, $\int h_i h_j(x)\,dx = \delta_{ij}$,

$$\begin{aligned} \theta(P) &= 1 + \sum_{i=1}^{\infty} \alpha_i^2 c_i^2 \\ &= 1 + \sum_{i=1}^{m-1} \alpha_i^2 c_i^2 + \sum_{i=m}^{\infty} \alpha_i^2 c_i^2. \end{aligned}$$

Let $\beta = (\alpha_1, \dots, \alpha_{k-1})$ and $\pi_{0\beta}$ be the conditional distribution of all the α's and Δ's given β. For any bounded loss function $L(\theta, a)$,

$$\inf_\delta E_{\pi_0} L(\theta, \delta) = \inf_\delta \int E_{\pi_{0\beta}} L(\theta, \delta)\nu(d\beta) \geq \int \inf_\delta E_{\pi_{0\beta}} L(\theta, \delta)\nu(d\beta), \tag{2.8}$$

where δ ranges over all estimates of θ based on $X_1, \dots, X_n$ and ν is the marginal distribution of β. Therefore, there exists a value β_0 of β such that the Bayes risk of π_0 is no smaller than the Bayes risk of $\pi_{0\beta_0} \equiv \pi_{00}$. Under π_{00}, if $m = [3\log_2 n]$ any interval of the form $[i2^{-m}, (i+1)2^{-m})$ contains at most one of $X_1, \dots, X_n$ with probability $\geq 1 - (2n)^{-1}$. Arguing as before, under π_{00}, except on a set of probability $O(n^{-1})$ the conditional distribution of $\Delta \equiv \{\Delta_k(i)\colon 1 \leq i \leq 2^k,\ k \geq m\}$ given $X_1, \dots, X_n$ is the same as the marginal distribution. We claim that the same is true of the conditional distribution of $\alpha = \{\alpha_k, \dots,\ k \geq m\}$. Write the joint density of $(\alpha, \Delta, X_1, \dots, X_n)$ with respect to the measure μ, where, under μ, the α_k's and $\Delta_k(i)$ have the distribution specified earlier and $X_1, \dots, X_n$ are independent of α, Δ and are uniform $(0,1)$ as

$$\prod_{i=1}^{n}\left(1 + \sum_{k=1}^{m-1} c_k \alpha_{k0} h_k(X_i) + \sum_{k=m}^{\infty} c_k \alpha_k h_k(X_i)\right).$$

The posterior density, if at most one X_i is in each interval $[i/2^k, (i+1)/2^k)$, $k \ge m$, is proportional to

$$\prod_{i=1}^{n}\left(A_i(X_i) + \sum_{k=m}^{\infty} c_k \alpha_k \varepsilon_k(X_i)\Delta_{ki}\right),$$

where $A_i(x) = 1 + \sum_{k=1}^{m-1} c_k \alpha_{k0} h_k(x)$, $\Delta_{ki} = \Delta_k(j)$ iff j is such that $X_i \in [j2^{-k}, (j+1)2^{-k})$ and

$$\varepsilon_k(X_i) = \begin{cases} +1, & \text{if } X_i \in [j2^{-k}, (j+\frac{1}{2})2^{-k}), \\ -1, & \text{if } X_i \in [(j+\frac{1}{2})2^{-k}, (j+1)2^{-k}). \end{cases}$$

Then the posterior probability that $(\alpha_{m+1}, \dots, \alpha_{m+t}) = (\alpha^0_{m+1}, \dots, \alpha^0_{m+t})$ given $X_1 = x_1, \dots, X_n = x_n$ is proportional to

$$E_\mu\left\{\prod_{i=1}^{n}\left(A_i(X_i) + \sum_{k=m+1}^{m+t} c_k \alpha^0_k \varepsilon_k(X_i)\Delta_{ki} + \sum_{k=m+t+1}^{\infty} c_k \alpha_k \varepsilon_k(X_i)\Delta_{ki}\right)\right.$$
$$\left.\times 1\left(\alpha_{m+1} = \alpha^0_{m+1}, \dots, \alpha_{m+t} = \alpha^0_{m+t}\right)\right\}.$$

But the α_k and the Δ_{ki} are independent under μ. Multiplying out the product and using the symmetry of the Δ_{ki}, we obtain that the posterior probability is proportional to $\prod_{i=1}^{n} A_i(X_i)$ and our claim follows. To complete the argument note that, under π_{00} if $B_m = \sum_{k=m}^{\infty} c_k^2(\alpha_k^2 - \frac{1}{2})$,

$$P\left[B_m \ge \tfrac{1}{2}c_m^2\right] \ge P\left[\alpha_m = 1, \sum_{k=m+1}^{\infty} c_k^2\left(\alpha_k^2 - \tfrac{1}{2}\right) \ge 0\right] \ge \tfrac{1}{4}$$

by the symmetry and independence of α_m, and $\alpha_k^2 - \frac{1}{2}$, $k = m+1, \dots$. A similar argument shows

$$P\left[B_m \le -\tfrac{1}{2}c_m^2\right] \ge \tfrac{1}{4}.$$

Hence, if at most one X_i falls in each interval,

$$\begin{aligned}
&\inf_a P\left[|\theta - a| \ge \tfrac{1}{2}c_m^2 | X_1, \dots, X_n\right] \\
&\quad \ge \min\left\{P\left[B_m \ge \tfrac{1}{2}c_m^2 | X_1, \dots, X_n\right], P\left[B_m \le -\tfrac{1}{2}c_m^2 | X_1, \dots, X_n\right]\right\} \\
&\quad \ge \tfrac{1}{4} + O_p(n^{-1}),
\end{aligned}$$

since, except on a set of probability $O(n^{-1})$, the marginal and conditional distributions of B_m agree. So the Bayes risk of π_{00} for the loss function

$L_m(\theta, a) = 1[|\theta - a| \geq \frac{1}{2}c_m^2]$ is $\geq \frac{1}{4} + O(n^{-1})$. If $c_m = 9\varepsilon^2[\log n]^{-1-\varepsilon}$, say, then (2.6) follows from (2.8). □

In the model of Engle, Granger, Rice and Weiss (1986) we observe $X_i = (W_i, Z_i, Y_i)$, $i = 1, \ldots, n$, where

$$Y = \beta W + t(Z) + \varepsilon \tag{2.9}$$

and $\varepsilon \sim \mathbf{N}(0, \sigma^2)$. The joint distribution of (W, Z) and t are unknown. In recent work, Chen (1988) and Cuzick (1987) have exhibited, under various smoothness restrictions on t, estimates $\hat{\beta}$ which are asymptotically $\mathbf{N}(0, I^{-1}/n)$, where

$$I = \sigma^{-2}E(W - E(W|Z))^2 > 0 \tag{2.10}$$

unless W is a function of Z. Local calculations yield this as the information bound whenever $W \in L_2$. Let

$\mathbf{P} = \{$All distributions (W, Z, Y) given by (2.9) such that $I > 0$ and well defined$\}$.

THEOREM 3. (1) *Even if* $\sigma = 0$ [*or, equivalently, I given by* (2.10) *equals* ∞], *there exists a subset* $\mathbf{P}_0$ *of* $\mathbf{P}$ *such that for all estimates* $\hat{\beta}_n$,

$$\sup_{\mathbf{P}_0} P\left[\left|\hat{\beta}_n - \beta\right| \geq \varepsilon\right] > 0 \quad \textit{for any } \varepsilon > 0. \tag{2.11}$$

(2) *For* $\sigma > 0$ *there exists a compact subset* $\mathbf{P}_0$ *of* $\mathbf{P}$ *such that for all estimates* $\hat{\beta}_n$ *and all* $\gamma > 0$,

$$\liminf_n \sup_{\mathbf{P}_0} \left[\left|\hat{\beta} - \beta\right| \geq n^{-\gamma}\right] > 0.$$

We argue as for Theorem 2.

PROOF OF THEOREM 3. (1) We give the simpler construction for $\sigma = 0$ and $\mathbf{P}_0$ noncompact and sketch if for $\sigma > 0$ and $\mathbf{P}_0$ compact. Here is the prior π_n. Take $W = \pm 1$ with probability $\frac{1}{2}$ and $0 \leq Z \leq 1$.

Let $\alpha, \Delta_0, \ldots, \Delta_m$, $m = n^3$, be i.i.d. and equal to ± 1 with probability $\frac{1}{2}$. If $\alpha = -1$, then $\beta = 0$, $Z \sim \mathbf{U}(0,1)$ independent of W and $t(z) \equiv 0$. If $\alpha = 1$, then $\beta = c$ and the conditional density of $Z|W$ and $t(\cdot)$ are given by

$$\begin{aligned} &p(z|w) = 1 - \Delta_i w, \qquad t(z) = c\Delta_i, \qquad \frac{i}{m+1} \leq z < \frac{i + 1/2}{m+1}, \\ &p(z|w) = 1 + \Delta_i w, \qquad t(z) = -c\Delta_i, \qquad \frac{i + 1/2}{m+1} \leq z < \frac{i+1}{m+1}. \end{aligned} \tag{2.12}$$

Again with probability $1 - O(n^{-1})$, the posterior of $\Delta_1, \ldots, \Delta_m$ is the same as the prior distribution. Note also by construction that $\beta W + t(Z) \equiv 0$. So, with probability $1 - O(n^{-1})$,

$$P[\alpha = 1 | W_i, Z_i, Y_i, i = 1, \dots, n] = P[\alpha = 1 | W_i, Z_i, i = 1, \dots, n]$$

is proportional to,

$$E\left\{ \prod_{i=1}^{n} (1 - \Delta_i W_i)^{\delta_i} (1 + \Delta_i W_i)^{1-\delta_i} \right\}, \tag{2.13}$$

where $W_1, Z_1, \dots, W_n, Z_n$ are fixed. If Z_i falls in $[j_i/(m+1), (j_i+1)/(m+1))$, we define $\delta_i = 1$ if Z_i is in the first half of that interval and 0 if it is in the second. The expectation in (2.13) is again 1 and we conclude that the posterior distribution of α is the same as its prior and hence that the Bayes risk of π_n is bounded away from 0. (2.11) follows.

(2) If $\sigma = 1$ (say), proceed as follows. Let $\alpha, \Delta_1, \dots, \Delta_m$ be as above. Suppose $P[W = 0] = P[W = 1] = \frac{1}{2}$ and that the conditional distribution of Z given $W = 0$ is $\mathbf{U}(0, 1)$. Under π_n if $\alpha = -1$, $\beta = 0$ and Z given $W = 1$ is also $\mathbf{U}(0, 1)$. Let

$$t_n(z) = \begin{cases} a_n \Delta_i, & i/(m+1) \le z < \left(i + \frac{1}{2}\right)/(m+1), \\ -a_n \Delta_i, & \left(i + \frac{1}{2}\right)/(m+1) \le z < i/(m+1). \end{cases}$$

If $\alpha = 1$, $\beta = c_n$ and

$$p(z | W = 1) = \begin{cases} 1 - b_n \Delta_i, & i/(m+1) \le z < \left(i + \frac{1}{2}\right)/(m+1), \\ 1 + b_n \Delta_i, & \left(i + \frac{1}{2}\right)/(m+1) \le z < (i+1)/(m+1). \end{cases} \tag{2.14}$$

With probability $1 - O(n^{-1})$, there is at most one Z_i in each interval $[i(m+1)^{-1}, (i+1)(m+1)^{-1})$. Conditional on that event, being given (W_i, Z_i, Y_i) is the same as being given (W_i, V_i, Y_i), where V_i is the fractional part of $(m+1)Z_i$. Further, the posterior distribution of β is the same as the conditional distribution of β given $\{(V_i, Y_i)\colon W_i = 1\}$. Given $W_i = 1$, V_i is $\mathbf{U}(0, 1)$ by (2.14) since the conditional distribution of Δ_{ji} given $W_i = 1$, where $Z_i \in (j_i/(m+1), (j_i+1)/(m+1))$, is the same as its prior.

Finally, the conditional density of Y_i given $W_i = 1$, V_i, $\alpha = 1$, is

$$\begin{aligned} &\tfrac{1}{2}(1 - b_n)\phi(y - c_n - a_n)' + \tfrac{1}{2}(1 + b_n)\phi(y - c_n + a_n) \\ &\quad = \phi(y) + y\phi(y)(c_n - a_n b_n) + O\left(c_n^2 + a_n^2\right). \end{aligned}$$

If $a_n = c_n^{1-\delta}$, $b_n = c_n^{\delta}$, $\delta > 0$, the density of Y_i given $W_i = 1$, V_i, $\alpha = 1$ is $\phi(y)(1 + c_n^{2-2\delta}h(y) + O(c_n^3 + a_n^3))$, where $\int \phi(y)h(y)\,dy = 0$. One can show the joint distribution of $\{(V_i, Y_i)\colon W_i = 1\}$ under $\alpha = 1$ is contiguous to that under $\alpha = 0$ provided $c_n^{2-2\delta} = O(n^{-1/2})$. Hence, by taking $c_n = n^{-1/4+\varepsilon}$, $\varepsilon > 0$, arbitrary, we can deduce that β cannot be estimated at a rate better than $n^{-1/4+\varepsilon}$. $\square$

3. Validity of the bounds for a class of models. We consider semiparametric models with the following structure:

$$\mathbf{P} = \bigcup_{m=1}^{\infty} \mathbf{P}_m, \qquad \mathbf{P}_m \subset \mathbf{P}_{m+1}, \forall\, m, \tag{3.1}$$

and $\mathbf{P}_m$ regular parametric. That is, we can write

$$\mathbf{P}_m = \{P_{(\theta,\eta^m)}: \theta \in \Theta,\ \eta^m = (\eta_1, \dots, \eta_{d-1}), \text{ with } d = d(m)$$
$$\text{and } \eta_j \in E_j,\ j = 1, \dots, d-1,\ E_j,\ \Theta \text{ open subsets of } R\}.$$

1. $\mathbf{P} \ll \mu$.
2. The maps $(\theta, \eta^m) \to P_{(\theta,\eta^m)}$ are 1–1 for all m. Further, if $P \in \mathbf{P}_m = \mathbf{P}_m \cap \mathbf{P}_{m'}$, $m' > m$, then the first $d(m)$ coordinates of $\eta^{m'}$ agree with η^m.
3. The maps $(\theta, \eta^m) \to s(\theta, \eta^m) \equiv (dP_{(\theta,\eta^m)}/d\mu)^{1/2} \in L_2(\mu)$ are continuously Fréchet differentiable with derivative $\dot{s}(\theta, \eta^m) = (\dot{s}_1, \dots, \dot{s}_d)(\theta, \eta^m)$, $\dot{s}_j \in L_2(\mu)$, $j = 1, \dots, d$.
4. The information matrix,

$$I(\theta, \eta^m) \equiv 4\left[\int \dot{s}_i \dot{s}_j(\theta, \eta^m)\, d\mu\right]_{d\times d} = \left[E_{(\theta,\eta^m)} \dot{l}_i \dot{l}_j(\theta, \eta^m)\right]_{d\times d},$$

is nonsingular for all (θ, η^m), where $\dot{l}(\theta, \eta^m) = 2(\dot{s}/s)(\theta, \eta^m)$ is the derivative of the log likelihood.

In words, every member of $\mathbf{P}$ belongs to a nice parametric model whose dimension d can, however, be arbitrarily large. A moment's thought will show that most if not all semiparametric models proposed in the literature can be thought of as the closures (for weak convergence) of such $\mathbf{P}$. For example, the symmetric location model $\{P$: P is absolutely continuous on R, symmetric about some $\theta \in R\}$ is the closure of $\mathbf{P}$ as in (3.1), where $P_{(\theta,\eta^m)}$, for example, has

$$\log p_{(\theta,\eta^m)}(x) = h(x - \theta, \eta^m),$$

where

$$h''(x, \eta^m) = \sum_{k=1}^{d-1} \eta_k 1(|x| < b_{km}),$$

where $d = 2^m + 1$, $b_{km} = mk2^{-m}$, $k = 1, \dots, d-1$. That is, we assume that the log density of $X - \theta$ is a symmetric quadratic spline with knots at $\pm b_{km}$, which is constant for $|x| > m$. Such models have been considered by Faraway (1987) and Stone (1986) among others. It is well known [see Le Cam (1956) and Bickel (1982)] that there exist estimates $\hat{\theta}_{mn}, \eta_{mn}$ which are efficient on $\mathbf{P}_m$. In particular,

$$\hat{\theta}_{mn} - \theta_0 = n^{-1} \sum_{i=1}^{n} \tilde{l}_{0m}(X_i) + o_{P_0}(n^{-1/2}), \tag{3.2}$$

where

$$\tilde{l}_{0m} = \frac{s^{-1}}{2} \frac{s_1^*}{\|s_1^*\|^2}$$

and

$$s_1^* = \dot{s}_1 - \Pi(\dot{s}_1 | [\dot{s}_2, \dots, \dot{s}_d]),$$

$\Pi(h|L)$ denotes the projection of $h \in L_2(\mu)$ on the closed linear subspace L in the $L_2(\mu)$ norm, $\|\cdot\|$, and $[\dot{s}_2, \dots, \dot{s}_d]$ is the linear span of $\{\dot{s}_2, \dots, \dot{s}_d\}$. $\hat{\eta}_{mn} - \eta_0$ has a similar expansion but we can only note that

$$\hat{\eta}_{mn} - \eta_0 = O_{P_0}(n^{-1/2}). \tag{3.3}$$

These relations hold for each m fixed, all $P_0 \in \mathbf{P}_m$, as $n \to \infty$. Frequently, we achieve (3.2) and (3.3) using the maximum likelihood estimates of θ, η^m under $\mathbf{P}_m$. For any $P \in \mathbf{P}$, let $\eta = (\eta_1, \dots, \eta_{d(P)})$, and $d(P)$ is the smallest m such that $P \in \mathbf{P}_m$. For the model $\mathbf{P}$, the information bound in estimating θ at $P_0 = P_{(\theta_0, \eta_0)}$ is given by

$$\| I^{-1}(P_0; \theta) = \tfrac{1}{4} \|\dot{s}_1 - \Pi \| (\dot{s}_1 | \dot{\zeta}_2(\theta_0, \eta_0)) \|^{-2},$$

where

$$\dot{\zeta}_2(\theta_0, \eta_0) = \text{closure of the linear span of } \{\dot{s}_2(\theta_0, \eta_0), \dots, \dot{s}_l(\theta_0, \eta_0), \dots\}.$$

Here, for $m \geq m(P_0)$, we consider P_0 as a member of $\mathbf{P}_m$, i.e., corresponding to (θ_0, η_0^m) such that $P_0 = P_{(\theta_0, \eta_0^m)}$.

Suppose $I(P_0; \theta) > 0$ for all $P_0 \in \mathbf{P}$. Let

$$\tilde{l}(\theta_0, \eta_0) = 2s^{-1}(\theta_0, \eta_0)(\dot{s}_1(\theta_0, \eta_0) - \Pi(\dot{s}_1(\theta_0, \eta_0) | \dot{\zeta}_2(\theta_0, \eta_0)) / I(P_0; \theta)) \tag{3.4}$$

be the efficient influence function for estimating θ in $\mathbf{P}$ at P_0; $\tilde{l}$ depends on (θ_0, η_0).

THEOREM 4. *Suppose that if* $P_{(\theta_k, \eta_k^m)} \in \mathbf{P}_m$, $\theta_k \to \theta_0$, $\eta_k^m \to \eta_0^m$, *then*

$$\Pi(v | \dot{\zeta}_2(\theta_k, \eta_k^m)) \to \Pi(v | \dot{\zeta}_2(\theta_0, \eta_0)) \tag{3.5}$$

for all $v \in L_2(\mu)$ *and*

$$\limsup_k \| \tilde{l}(\theta_k, \eta_k^m) \|_\infty < \infty, \tag{3.6}$$

where $\|\cdot\|_\infty$ *is the* sup *norm.*

Then there exists $\hat{\theta}_n$ *such that,*

$$\hat{\theta}_n = \theta_0 + n^{-1} \sum_{i=1}^n \tilde{l}_0(X_i) + o_{P_0}(n^{-1/2}),$$

where $\tilde{l} = \tilde{l}(\theta_0, \eta_0)$.

Moreover, the $\hat{\theta}_n$ are at least locally regular. That is, for all $P_0 \in \mathbf{P}$, $\{P_\tau : |\tau| < 1\}$ is a regular parametric submodel of $\mathbf{P}$, $\tau_n = O(n^{-1/2})$, we have $\mathbf{L}_{\tau_n}(n^{1/2}(\hat{\theta}_n - \theta(P_{\tau_n})))$ tending to a limit law independent of $\{P_\eta\}$.

The construction is essentially to pick the lowest dimensional submodel $\mathbf{P}_{\hat{m}_n}$ which is close enough to the empirical distribution, then treat $\hat{m}_n$ as fixed, compute the efficient estimate $\hat{\eta}_{\hat{m}_n n}$ of $\eta_{\hat{m}_n}$ in that model and then "solve the equation;"

$$\sum_{i=1}^{n} \tilde{l}(\theta, \hat{\eta}_{m_n n}) = 0. \tag{3.7}$$

The resulting estimate is well behaved if $P \in \mathbf{P}$. However, if $P \in \overline{\mathbf{P}} - \mathbf{P}$, we necessarily have $\hat{m}_n \to \infty$ and no guarantee that the solution of (3.7) is even consistent, much less efficient. In fact, the examples of the previous section make it clear that there is no hope for such a general consistency theorem. The question remains whether one can formulate reasonable conditions on the structure of $\tilde{l}$ and the behaviour of the distance in suitable metrics $\mathbf{P}_m$ and members of $\overline{\mathbf{P}} - \mathbf{P}$ as a function of m which yield the validity of the information bounds for members of $\mathbf{P}$. An attempt in this direction is the work of Severini and Wong (1987). However, we do not pursue this, in part, because we believe that the checking of any such conditions in models of interest will be at least as difficult as the construction of efficient estimates by one of a number of heuristic methods which have been developed—see BKRW, Chapter 7 for a discussion.

Proof. Let d_K be the Kolmogorov distance between distributions. Let $\hat{\theta}_{mn}, \hat{\eta}_{mn}$ be as in (3.2) and (3.3) and let

$$\hat{P}_m \text{ be the corresponding member of } \mathbf{P}_m.$$

Let $\hat{m}_n$ be the first m such that $d_K(\hat{P}_m, P_n) \le \varepsilon_n$, where $\varepsilon_n \to 0$, $n^{1/2}\varepsilon_n \to \infty$, P_n is the empirical distribution. Evidently, if $m_0 = m(P_{(\theta_0, \eta_0)})$,

$$P_0[\hat{m}_n = m_0] \to 1.$$

Moreover, $\hat{P}_{\hat{m}_n} \leftrightarrow (\hat{\theta}_{\hat{m}_n n}, \hat{\eta}_{m_n n}) = (\theta_0, \eta_0) + O_{p_0}(n^{-1/2})$. Therefore, by (3.5),

$$\int \left(\tilde{l}(\theta_n, \hat{\eta}_{m_n n}) - \tilde{l}(\theta_n, \eta_n) \right)^2 s^2(\theta_n, \eta_n)\, d\mu = o_{p_0}(1), \tag{3.8}$$

for all sequences $P_{(\theta_n, \eta_n)} \in \mathbf{P}_{m_0}$ with $|\theta_n - \theta_0| = O(n^{-1/2})$, $|\eta_n - \eta_0| = O(n^{-1/2})$.

Moreover, using (3.6), we see that,

$$\begin{aligned}
&\int \tilde{l}(\theta_n, \hat{\eta}_{\hat{m}_n n}) s^2(\theta_n, \eta_n)\, d\mu \\
&\quad = 2\int \tilde{l}(\theta_n \hat{\eta}_{m_0 n})\big(s(\theta_n, \eta_n) - s(\theta_n, \hat{\eta}_{m_0 n})\big) s(\hat{\theta}_n, \hat{\eta}_{m_0 n})\, d\mu \\
&\qquad + O_{p_0}\left(\| s(\theta_n, \eta_n) - s(\theta_n, \hat{\eta}_{m_0 n}) \|^2 \right) \\
&\quad = 2\int \tilde{l}(\theta_n, \hat{\eta}_{m_0 n})\big(\dot{s}_2(\theta_n, \hat{\eta}_{m_0 n}), \dots, \dot{s}_{m_0}(\theta_m, \hat{\eta}_{m_0 n})\big) \\
&\qquad \times (\eta_n - \hat{\eta}_{m_0 n})' s(\hat{\theta}_n, \hat{\eta}_{m_0 n})\, d\mu \\
&\qquad + o_{p_0}\left(|\eta_n - \hat{\eta}_{m_0 n}| \right) + O_{p_0}\left(\| s(\theta_n, \eta_n) - s(\theta_n, \hat{\eta}_{m_0 n}) \|^2 \right).
\end{aligned} \tag{3.9}$$

The first term on the right in (3.9) is 0 by (3.4). The last two terms are

$o_{P_0}(n^{-1/2})$ by (3.2) and (3.3), so

$$\int \tilde{l}(\theta_n, \hat{\eta}_{\hat{m}_n n}) s^2(\theta_n, \eta_n)\, du = o_{p_0}(n^{-1/2}). \tag{3.10}$$

Together, (3.8) and (3.10) yield the existence of $\hat{\theta}_n$—see Klassen (1987), for example. □

Thus the $\hat{\theta}_n$ are at least locally regular and $n^{1/2}(\hat{\theta}_n - \theta_0)$ is asymptotically normal $(0, I^{-1}(P_0;\theta))$, i.e., achieves the information bound.

NOTE. (1) Conditions (3.5) and (3.6) are trivially satisfied by the symmetric location example. Condition (3.6) can be interpreted as a robustness condition for efficient estimates in $\mathbf{P}_m$. That is, on the model $\mathbf{P}_m$, efficient influence functions are bounded and bounded uniformly in small Hellinger neighbourhoods of any P.

(2) It is easy to check that if in the model of Engle, Granger, Rice and Weiss we, for instance, let $\mathbf{P}_m$ be such that $t(Z)$ and $\log P(W=1|Z)$ are representable as splines with $d(m)$ knots, condition (3.5) is satisfied. Although condition (3.6) fails for ε Gaussian, $\tilde{l}$ is of the form ε times functions which are uniformly $\|\cdot\|_\infty$ bounded and (3.7) continues to hold.

(3) A further peculiarity of these models is that, if we only consider the asymptotic behaviour of $\hat{\theta}_n$ at fixed (θ, η), it is asymptotically inadmissible. However, when we consider its behaviour over "contiguous" neighbourhoods in $\mathbf{P}$, it is uniquely asymptotically minimax. More precisely, let $\{P_t, |t|<1\}$ be a regular parametric submodel of $\mathbf{P}$ passing through $P_0 = P_{(\theta_0, \eta_0)}$. Corresponding to this model is its score function at (θ_0, η_0) given by (say) $s_0^{-1}v$, where $v \in \dot{\zeta}_2(\theta_0, \eta_0)$. Consider $\hat{\hat{\theta}} \equiv \hat{\theta}_{\hat{m}_n n}$. By Le Cam's third lemma, if $\theta_n \equiv \theta_n(t) = \theta(P_{tn^{-1/2}})$, $\eta_n \equiv \eta_n(t) = \eta(P_{tn^{-1/2}})$, then

$$L_{(\theta_n, \eta_n)}\sqrt{n}\left(\hat{\hat{\theta}} - \theta_n\right) \to \mathbf{N}\left(2t\int v s_1^*\, d\mu, \tfrac{1}{4}\|s_1^*\|^{-2}\right). \tag{3.11}$$

On the other hand, by the same argument,

$$L_{(\theta_n, \eta_n)}\sqrt{n}\left(\hat{\theta} - \theta_n\right) \to \mathbf{N}\left(0, I^{-1}(P_0;\theta)\right).$$

Now,

$$\begin{aligned} I(P_0;\theta) &= \tfrac{1}{4}\left\|\dot{s}_1 - \Pi\left(\dot{s}_1 | \dot{\zeta}_2(\theta_0, \eta_0)\right)\right\|^2 \\ &\le \frac{\|s_1^*\|^2}{4}. \end{aligned}$$

So, at (θ_0, η_0), i.e., $t=0$, both $\sqrt{n}(\hat{\theta} - \theta_0)$ and $\sqrt{n}(\hat{\hat{\theta}} - \theta_0)$ are asymptotically normal with mean 0 and the asymptotic variance of $\sqrt{n}\,\hat{\hat{\theta}}$ is smaller than that of $\hat{\theta}$. However, evidently, on each parametric submodel, for any bounded bowl-shaped loss function l,

$$\liminf_M \liminf_n \sup\left\{E_{(\theta_n(t), \eta_n(t))} l\left(n^{1/2}\left(\hat{\hat{\theta}} - \theta_n\right)\right) : |t| \le M n^{-1/2}\right\} = \sup_d l(d),$$

higher than the comparable asymptotic minimax risk for $\hat{\theta}$.

This is a superefficiency phenomenon. The estimator $\hat{\hat{\theta}}$ is, in view of (3.11), not locally regular, i.e., the limit of $\mathbf{L}_{(\theta_n, \eta_n)}(\sqrt{n}(\hat{\hat{\theta}} - \theta_n))$ is not independent of t.

REFERENCES

BEGUN, J. M., HALL. W. J., HUANG, W. M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.* **11** 432–452.

BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671.

BICKEL, P. J. and RITOV, J. (1988). Estimating integrated squared derivatives. *Sankhyā*. To appear.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1989). Efficient and Adaptive Inference in Semiparametric Models. Forthcoming monograph, Johns Hopkins University Press, Baltimore.

BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densités: Risque minimax. *Z. Warsch. Verw. Gebiete* **47** 119–137.

CHEN, H. (1988). Convergence rates for the parametric component in a partially linear model. *Ann. Statist.* **16** 136–146.

CUZICK, J. (1987). Semiparametric additive regression. Technical Report, Imperial Cancer Research Laboratories, London.

DONOHO, D. L. and LIU, R. C. (1988). Geometrizing rates of convergence. Technical Report, Dept. Statist., Univ. California, Berkeley.

ENGLE, R. F., GRANGER, C. W. J., RICE, J. and WEISS, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81** 310–320.

FARAWAY, J. J. (1987). Smoothing in adaptive estimation. Ph.D. dissertation, Univ. California, Berkeley.

HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berkeley Symp. Math Statist. Prob.* **1** 175–194. Univ. California Press.

HASMINSKII, R. and IBRAGIMOV, I. A. (1979). On the nonparametric estimation of functionals. In *Proc. 2nd Prague Symp. Asymptotic Statist.* (P. Mandl and M. Huskova, eds.) 41–51. North-Holland, Amsterdam.

KLAASSEN, C. A. J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.* **15** 1548–1562.

LE CAM, L. (1956). On the asymptotic theory of estimation and testing hypotheses. *Proc. Third Berkeley Symp. Math Statist. Prob.* **1** 129–156. Univ. California Press.

LEVIT, B. Y. (1978). Infinite dimensional information bounds. *Theor. Probab. Appl.* **20** 723–740.

PFANZAGL, J. (1982). *Contributions to a General Asymptotic Statistical Theory. Lecture Notes in Statist.* **13**. Springer, New York.

SCHWEDER, T. (1975). Window estimation of the asymptotic variance of rank estimation of location. *Scand. J. Statist.* **2** 113–126.

SEVERINI, T. A. and WONG, W.-H. (1987). Profile likelihood and semiparametric models. Technical Report, Univ. Chicago.

STEIN, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 187–195. Univ. California Press.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

STONE, C. J. (1986). A nonparametric framework for statistical modeling. Technical Report, Dept. Statist., Univ. California, Berkeley.

DEPARTMENT OF STATISTICS
THE HEBREW UNIVERSITY OF JERUSALEM
JERUSALEM 91905
ISRAEL

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

Chapter 6
Boostrap Resampling

Peter Hall

6.1 Introduction to Four Bootstrap Papers

6.1.1 Introduction and Summary

In this short article we discuss four of Peter Bickel's seminal papers on theory and methodology for the bootstrap. We address the context of the work as well as its contributions and influence. The work began at the dawn of research on Efron's bootstrap. In fact, Bickel and his co-authors were often the first to lay down the directions that others would follow when attempting to discover the strengths, and occasional weaknesses, of bootstrap methods.

Peter Bickel made major contributions to the development of bootstrap methods, particularly by delineating the range of circumstances where the bootstrap is effective. That topic is addressed in the first, second and fourth papers treated here. Looking back over this work, much of it done 25–30 years ago, it quickly becomes clear just how effectively these papers defined the most appropriate directions for future research.

We shall discuss the papers in chronological order, and pay particular attention to the contributions made by Bickel and Freedman (1981), since this was the first article to demonstrate the effectiveness of bootstrap methods in many cases, as well as to raise concerns about them in other situations. The results that we shall introduce in Sect. 6.1.2, when considering the work of Bickel and Freedman (1981), will be used frequently in later sections, especially Sect. 6.1.5.

The paper by Bickel and Freedman (1984), which we shall discuss in Sect. 6.1.3, pointed to challenges experienced by the bootstrap in the context of stratified

P. Hall (✉)
Department of Mathematics and Statistics, The University of Melbourne,
Melbourne, VIC, Australia
e-mail: halpstat@ms.unimelb.edu.au

J. Fan et al. (eds.), *Selected Works of Peter J. Bickel*, Selected Works in Probability and Statistics 13, DOI 10.1007/978-1-4614-5544-8_6,

sampling. This is ironic, not least because some of the earliest developments of what, today, are called bootstrap methods, involved sampling problems; see, for example, Jones (1956), Shiue (1960), Gurney (1963) and McCarthy (1966, 1969).

Section 6.1.4 will treat the work of Bickel and Yahav (1988), which contributed very significantly to methodology for efficient simulation, at a time when the interest in this area was particularly high. Bickel et al. (1997), which we shall discuss in Sect. 6.1.5, developed deep and widely applicable theory for the m-out-of-n bootstrap. The authors showed that their approach overcame consistency problems inherent in the conventional n-out-of-n bootstrap, and gave rates of convergence applicable to a large class of problems.

6.1.2 Laying Foundations for the Bootstrap

Thirty years ago, when Efron's (1979) bootstrap method was in its infancy, there was considerable interest in the extent to which it successfully accomplished its goal of estimating parameters, variances, distributions etc. As Bickel and Freedman (1981) noted, Efron's paper "gives a series of examples in which [the bootstrap] principle works, and establishes the validity of the approach for a general class of statistics when the sample space is finite." Bickel and Freedman (1981) set out to assess the bootstrap's success in a much broader setting than this.

In the early 1980s, saying that the bootstrap "works" meant that bootstrap methods gave consistent estimators, and in this sense were competitive with more conventional methods, for example those based on asymptotic analysis. Within about 5 years the goals had changed; it had been established that bootstrap methods "work" in a very wide variety of circumstances, and, although there were counterexamples to this general rule, by the mid 1980s the task had become largely one of comparing the effectiveness of the bootstrap relative to more conventional techniques. But in 1981 the extent to which the bootstrap was consistent was still largely unknown. Bickel and Freedman (1981) contributed mightily to the process of discovery there.

In particular, Bickel and Freedman (1981) were the first to establish rigorously that bootstrap methodology is consistent in a wide range of settings. The impact of their paper was dramatic. It provided motivation for exploring the bootstrap more deeply in a great many settings, and furnished some of the mathematical tools for that development. In the same year, in fact in the preceding paper in the *Annals*, Singh (1981) explored second-order properties of the bootstrap. However, Bickel and Freedman (1980) also took up that challenge at a particularly early stage.

As a prelude to describing the results of Bickel and Freedman (1981) we give some notation. Let $\chi_n = X_1, \dots, X_n$ denote a sample of n independent observations from a given univariate distribution with finite variance σ^2, write $\bar{X}_n = n^{-1} \sum_i X_i$ for the sample mean, and define

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 ,$$

the bootstrap estimator of σ^2. Let $\chi_m^* = \{X_1^*, \ldots, X_m^*\}$ denote a resample of size m drawn by sampling randomly, with replacement, from χ, and put $\bar{X}_m^* = m^{-1} \sum_{i \le m} X_i^*$. Bickel and Freedman's (1981) first result was that, in the case of m-resamples, the m-resample bootstrap version of $\hat{\sigma}_n^2$, i.e.

$$\hat{\sigma}_m^{*2} = \frac{1}{m} \sum_{i=1}^{m} (X_i^* - \bar{X}_m^*)^2 ,$$

converges to σ^2 as both m and n increase, in the sense that, for each $\varepsilon > 0$,

$$P(|\hat{\sigma}_m^* - \sigma| > \varepsilon \,|\, \chi_n) \to 0 \tag{6.1}$$

with probability 1. Moreover, Bickel and Freedman (1981) showed that the conditional distribution of $m^{1/2}\,(\bar{X}_m^* - \bar{X}_n)$, given χ_n, converges to the normal $\mathrm{N}(0, \sigma^2)$ distribution. Taking $m = n$, the latter property can be restated as follows:

$$\begin{aligned}&\text{the probabilities } P\{n^{1/2}\,(\hat{\theta}^* - \hat{\theta}) \le \sigma\, x \,|\, \chi_n\} \text{ and } P\{n^{1/2}\,(\hat{\theta} - \theta) \le \sigma\, x\}\\ &\text{both converge to } \Phi(x)\text{, the former converging with probability 1,}\end{aligned} \tag{6.2}$$

where Φ denotes the standard normal distribution and, on the present occasion, $\theta = E(X_i)$, $\hat{\theta} = \bar{X}_n$ and $\hat{\theta}^* = \bar{X}_n^*$.

The second result established by Bickel and Freedman (1981) was a generalisation of this property to multivariate settings. Highlights of subsequent parts of the paper included its contributions to theory for the bootstrap in the context of functionals of a distribution function. For example, Bickel and Freedman (1981) considered von Mises functionals of a distribution function H, defined by

$$g(H) = \int\!\!\int \omega(x,y)\, dH(x)\, dH(y) ,$$

where the function ω of two variables is symmetric, in the sense that $\omega(x,y) = \omega(y,x)$, and where

$$\int\!\!\int \omega(x,y)^2\, dH(x)\, dH(y) + \int \omega(x,x)^2\, dH(x) < \infty . \tag{6.3}$$

If we take H to be either $\widehat{F}_n$, the empirical distribution function of the sample χ_n, or $\widehat{F}_n^*$, the version of $\widehat{F}_n$ computed from χ_n^*, then

$$g(\widehat{F}_n) = \frac{1}{n^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \omega(X_{i_1}, X_{i_2}) , \quad g(\widehat{F}_n^*) = \frac{1}{n^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \omega(X_{i_1}^*, X_{i_2}^*) .$$

Bickel and Freedman (1981) studied properties of this quantity. In particular they proved that if (6.3) holds with $H = F$, denoting the common distribution function of the X_is, then the distribution of $n^{1/2}\{g(\widehat{F}_n^*) - g(\widehat{F}_n)\}$, conditional on the data, is asymptotically normal $\mathrm{N}(0, \tau^2)$ where

$$\tau^2 = 4\left[\int \left\{\int \omega(x,y)\, dF(y)\right\}^2 dF(x) - g(F)^2\right].$$

This limit distribution is the same as that of $n^{1/2}\{g(\widehat{F}_n) - g(F)\}$, and so the above result of Bickel and Freedman (1981) confirms, in the context of von Mises functions of the empirical distribution function, that (6.2) holds once again, provided that σ there is replaced by τ and we redefine $\theta = g(F)$, $\hat{\theta} = g(\widehat{F}_n)$ and $\hat{\theta}_n^* = g(\widehat{F}_n^*)$. That is, the bootstrap correctly captures, once more, first-order asymptotic properties. Subsequent results of Bickel and Freedman (1981) also showed that the same property holds for the empirical process, and in particular that the process $n^{1/2}(\widehat{F}_n^* - \widehat{F}_n)$ has the same first-order asymptotic properties as $n^{1/2}(\widehat{F}_n - F)$. Bickel and Freedman (1981) also derived the analogue of this result for the quantile process.

Importantly, Bickel and Freedman (1981) addressed cases where the bootstrap fails to enjoy properties such as (6.2). In their Sect. 6 they gave two counterexamples, one involving U-statistics and the other, spacings between extreme order statistics, where the bootstrap fails to capture large-sample properties even to first order. In both settings the problems are attributable, at least in part, to failure of the bootstrap to correctly capture the relationships among very high-ranked, or very low-ranked, order statistics, and in that context we shall relate below some of the issues to which Bickel and Freedman's (1981) work pointed. This account will be given in detail because it is relevant to later sections.

Let $X_{(1)} < \ldots < X_{(n)}$ denote the ordered values in χ_n; we assume that the common distribution of the X_is is continuous, so that the probability of a tie equals zero. In this case the probability, conditional on χ_n, of the event ε_n that the largest X_i, i.e. $X_{(n)}$, is in χ_n^*, equals 1 minus the conditional probability that $X_{(n)}$ is not contained in in χ_n^*. That is, it equals $1 - (1 - n^{-1})^n = 1 - e^{-1} + O(n^{-1})$. Therefore, as $n \to \infty$,

$$P(X_{(n)}^* = X_{(n)} \,|\, \chi_n) = P(X_{(n)} \in \chi_n^* \,|\, \chi_n) \to 1 - e^{-1},$$

where the convergence is deterministic. Similarly, for each integer $k \geq 1$,

$$\pi_{nk} \equiv P(X_{(n)}^* = X_{(n-k)} \,|\, \chi_n) \to \pi_k \equiv e^{-k}\left(1 - e^{-1}\right) \tag{6.4}$$

as $n \to \infty$; again the convergence is deterministic. Consequently the distribution of $X_{(n)}^*$, conditional on χ_n, is a mixture, and in particular is equal to $X_{(n-k)}$ with probability π_{nk}, for $k \geq 1$. Therefore:

given $\varepsilon>0$ and any metric, for example the Lévy metric, between distributions, we may choose $k = k(\varepsilon) \geq 1$ so large that the distribution of $X^*_{(n)}$, conditional on χ_n, is no more than ε from the discrete mixture $\sum_{0\leq j\leq k} X_{(n-j)}\, I_j$,where (a) exactly one of the random variables $I_1, I_2, \ldots$ is nonzero, (b) that variable takes the value 1, and (c) $P(I_k = 1) = \pi_k$ for $k \geq 0$. The upper bound of ε applies deterministically, in that it is valid with probability 1, in an unconditional sense. (6.5)

To indicate the implications of this property we note that, for many distributions F, there exist constants a_n and b_n, at least one of them diverging to infinity in absolute value as n increases; and a nonstationary stochastic process $\xi_1, \xi_2, \ldots$; such that, for each $k \geq 0$, the joint distribution of $(X_{(n)} - a_n)/b_n, \ldots, (X_{(n-k)} - a_n)/b_n$ converges to the distribution of $(\xi_1, \ldots, \xi_k)$. See, for example, Hall (1978). In view of (6.5) the distribution function of $(X^*_{(n)} - a_n)/b_n$, conditional on χ_n, converges to that of

$$Z = \sum_{j=0}^{\infty} \xi_j\, I_j\,,$$

where the sequence $I_1, I_2, \ldots$ is distributed as in (6.5) and is chosen to be independent of $\xi_1, \xi_2, \ldots$. In this notation,

$$P(X^*_{(n)} - a_n \leq b_n\, z \,|\, \chi_n) \to P(Z \leq z) \tag{6.6}$$

in probability, whenever z is a continuity point of the distribution of Z. On the other hand,

$$P(X_{(n)} - a_n \leq b_n\, z) \to P(\xi_1 \leq z)\,. \tag{6.7}$$

A comparison of (6.6) and (6.7) reveals that there is little opportunity for estimating consistently the distribution of $X_{(n)}$, using standard bootstrap methods. Bickel and Freedman (1981) first drew our attention to this failing of the conventional bootstrap. The issue was to be the object of considerable research for many years after the appearance of Bickel and Freedman's paper. Methodology for solving the problem, and ensuring consistency, was eventually developed and scrutinised; commonly the m-out-of-n bootstrap is used. See, for example, Swanepoel (1986), Bickel et al. (1997) and Bickel and Sakov (2008).

6.1.3 *The Bootstrap in Stratified Sampling*

Bickel and Freedman (1984) explored properties of the bootstrap in the case of stratified sampling from finite or infinite populations, and concluded that, with appropriate scaling, the bootstrap can give consistent distribution estimators in cases where asymptotic methods fail. However, without the proper scaling the bootstrap can be inconsistent.

The problem treated is that of estimating a linear combination,

$$\gamma = \sum_{j=1}^{p} c_j \mu_j \,, \tag{6.8}$$

of the means $\mu_1, \ldots, \mu_p$ of p populations $\Pi_1, \ldots, \Pi_p$ with corresponding distributions $F_1, \ldots, F_p$. The c_js are assumed known, and the μ_js are estimated from data. To construct estimators, a random sample $\chi(j) = \{X_{j1}, \ldots, X_{jn_j}\}$ is drawn from the jth population, and the sample mean $\underline{X}(j) = n_j^{-1} \sum_i X_{ji}$ is computed in each case. Bickel and Freedman (1984) considered two different choices of c_j, valid in two respective cases: (a) if it is known that each $E(X_{ji}) = \mu$, not depending on j, and that the variance σ_j^2 of Π_j is proportional to r_j, say, then

$$c_j = \frac{n_j / r_j}{\sum_k (n_k / r_k)} \,;$$

and (b) if the populations are finite, and in particular Π_j is of size N_j for $j = 1, \ldots, p$, then

$$c_j = \frac{N_j}{\sum_k N_k} \,.$$

In either case the estimator $\hat{\gamma}$ of γ reflects the definition of γ at (6.8):

$$\hat{\gamma} = \sum_{j=1}^{p} c_j \bar{X}(j) \,,$$

where $\bar{X}(j)$ is the mean value of the data in $\chi(j)$.

In both cases Bickel and Freedman (1984) showed that, particularly if the sample sizes n_j are small, the bootstrap estimator of the distribution of $\hat{\gamma} - \gamma$ is not necessarily consistent, in the sense that the distribution estimator minus the true distribution may not converge to zero in probability. The asymptotic distribution of $\hat{\gamma} - \gamma$ is normal $\mathrm{N}(0, \tau_1^2)$, say; and the bootstrap estimator of that distribution, conditional on the data, is asymptotically normal $\mathrm{N}(0, \tau_2^2)$; but the ratio τ_1^2 / τ_2^2 does not always converge to 1. Bickel and Freedman (1984) demonstrated that this difficulty can be overcome by estimating scale externally to the bootstrap process, in effect incorporating a scale correction to set the bootstrap on the right path. Bickel and Freedman also suggested other, more ad hoc remedies.

These contributions added immeasurably to our knowledge of the bootstrap. Combined with the counterexamples given earlier by Bickel and Freedman (1981), those authors showed that the bootstrap was not a device that could be used naively in all cases, without careful consideration.

Some researchers, a little outside the statistics community, had felt that bootstrap resampling methods freed statisticians from influence by a mathematical “priesthood” which was “frank about viewing resampling as a frontal attack upon their own situations” (Simon 1992). To the contrary, the work of Bickel and Freedman (1981,

1984) showed that a mathematical understanding of the problem was fundamental to determining when, and how, to apply bootstrap methods successfully. They demonstrated that mathematical theory was able to provide considerable assistance to the introduction and development of practical bootstrap methods, and they provided that aid to statisticians and non-statisticians alike.

6.1.4 Efficient Bootstrap Simulation

By the mid to late 1980s the strengths and weaknesses of bootstrap methods were becoming more clear, especially the strengths. However, computers with power comparable to that of today's machines were not readily available at the time, and so efficient methods were required for computation. The work of Bickel and Yahav (1988) was an important contribution to that technology. It shared the limelight with other approaches to achieving computational efficiency, including the balanced bootstrap, which was a version for the bootstrap of Latin hypercube sampling and was proposed by Davison et al. (1986) (see also Graham et al. 1990); importance resampling, suggested by Davison (1988) and Johns (1988); the centring method, proposed by Efron (1990); and antithetic resampling, introduced by Hall (1990).

The main impediment to quick calculation for the bootstrap was the resampling step. In the 1980s, when for many of us computing power was in short supply, bootstrap practitioners nevertheless advocated thousands, rather than hundreds, of simulations for each sample. For example Efron (1988), writing for an audience of psychologists, argued that "It is not excessive to use 2,000 replications, as in this paper, though we might have stopped at 1,000." In fact, if the number of simulations, B, is chosen so that the nominal coverage level of a confidence interval can be expressed as $b/(B+1)$, where b is an integer, then the size of B has very little bearing on the coverage accuracy of the interval; (see Hall 1986). However, choosing B too small can result in overly variable Monte Carlo approximations to endpoints for bootstrap confidence intervals, and to critical points for bootstrap hypothesis tests.

It is instructive here to relate a story that G.S. Watson told me in 1988, the year in which Bickel and Yahav's paper was published. Throughout his professional life Watson was an enthusiast of the latest statistical methods, and the bootstrap was no exception. Shortly after the appearance of Efron's (1979) seminal paper he began to experiment with the percentile bootstrap technique. Not for Watson a tame problem involving a sample of scalar data; in what must have been one of the first applications of the bootstrap to spatial or spherical data, he used that technique to construct confidence regions for the mean direction derived from a sample of points on a sphere. He wrote a program that constructed bootstrap confidence regions, put the code onto a floppy disc, and passed the disc to a Princeton geophysicist to experiment with. This, he told the geophysicist, was the modern alternative to conventional confidence regions based on the von Mises-Fisher distribution. The latter regions, of course, took their shape from the mathematical form of the fitted distribution, with relatively little regard for any advice that the data might have to offer. What did the geophysicist think of the new approach?

In due course Watson received a reply, to the effect that the method was very interesting and remarkably flexible, adapting itself well to quite different datasets. But it had a basic flaw, the geophysicist said, that made it unattractive—every time he applied the code on the floppy disc to the same set of spherical data, he got a different answer! Watson, limited by the computational resources of the day, and by the relative complexity of computations on a sphere, had produced software that did only about $B = 40$ simulations each time the algorithm was implemented. Particularly with the extra degree of freedom that two dimensions provided for fluctuations, the results varied rather noticeably from one time-based simulation seed to another.

This tale defines the context of Bickel and Yahav's (1988) paper. Their goal was to develop algorithms for reducing the variability, and enhancing the accuracy in that sense, of Monte Carlo procedures for implementing the bootstrap. Their approach, a modification for the bootstrap of the technique of Richardson extrapolation (a classical tool in numerical analysis; see Jeffreys and Jeffreys 1988, p. 288), ran as follows. Let $\widehat{F}_n$ (not to be confused with the same notation, but having a different meaning, in Sect. 6.1.2) denote the data-based distribution function of interest, and let F_n be the quantity of which $\widehat{F}_n$ is an approximation. For example, $\widehat{F}_n(x)$ might equal $P(\hat{\theta}_n^* - \hat{\theta}_n \leq x \,|\, \chi_n)$, where $\hat{\theta}_n$ denotes an estimator of a parameter θ, computed from a random sample χ_n of size n, in which case $\hat{\theta}_n^*$ would be the bootstrap version of $\hat{\theta}_n$. (In this example, $F_n(x) = P(\hat{\theta}_n - \theta \leq x)$.) Instead of estimating $\widehat{F}_n$ directly, compute estimators of the distribution functions $\widehat{F}_{n_1}, \ldots, \widehat{F}_{n_r}$, where the sample sizes $n_1, \ldots, n_r$ are all smaller than n, and in fact so small that $n_1 + \ldots + n_r$ is markedly less than n. In some instances we may also know the limit F_∞ of F_n, or at least its form, $\widetilde{F}_\infty$ say, constructed by replacing any unknown quantities (for example, a variance) by estimators computed from χ_n. The quantities $\widehat{F}_{n_1}, \ldots, \widehat{F}_{n_r}$ and $\widetilde{F}_\infty$ are much less expensive, i.e. much faster, to compute than $\widehat{F}_n$, and so, by suitable "interpolation" from these functions, we can hope to get a very good approximation to $\widehat{F}_n$ without going to the expense of actually calculating the latter.

In general the cost of simulating, or equivalently the time taken to simulate, is approximately proportional to $C_n B$, where C_n depends only on n and increases with that quantity. Techniques for enhancing the performance of Monte Carlo methods can either directly produce greater accuracy for a given value of B (the balanced bootstrap has this property), or reduce the value of C_n and thereby allow a larger value of B (hence, greater accuracy from the viewpoint of reduced variability) for a given cost. Bickel and Yahav's (1988) method is of the latter type. By enabling a larger value of B it alleviates the problem encountered by Watson and his geophysicist friend.

Bickel and Yahav's (1988) technique is particularly widely applicable, and has the potential to improve efficiency more substantially than, say, the balanced bootstrap. Today, however, statisticians' demands for efficient bootstrap methods have been largely assuaged by the development of more powerful computers. In the last 15 years there have been very few new simulation algorithms tailored to the bootstrap. Philippe Toint's aphorism that "I would rather have today's algorithms on yesterday's computers, than vice versa," loses impact when an algorithm is to some extent problem-specific, and its implementation requires skills that go beyond those needed to purchase a new, faster computer.

6.1.5 The m-Out-of-n Bootstrap

The m-out-of-n bootstrap is another example revealing that, in science, less is often more. Bickel and Freedman (1981, 1984) had shown that the standard bootstrap can fail, even at the level of statistical consistency, in a variety of settings; and, as we noted in Sect. 6.1.2, the m-out-of-n bootstrap, where m is an order of magnitude smaller than n, is often a remedy. Swanepoel (1986) was the first to suggest this method, which we shall define in the next paragraph. Bickel et al. (1997) made major contributions to the study of its theoretical properties. We shall give an example that provides further detail than we gave in Sect. 6.1.2 about the failure of the bootstrap in certain cases. Then we shall summarise briefly the contributions made by Bickel et al. (1997).

Consider drawing a resample $\chi_m^* = \{X_1^*, \ldots, X_m^*\}$, of size m, from the original dataset $\chi_n = \{X_1, \ldots, X_n\}$ of size n, and let $\hat{\theta} = \hat{\theta}_n$ denote the bootstrap estimator of θ computed from χ_n. In particular, if we can express θ as a functional, say $\theta(F)$, of the distribution function F of the data X_i, then

$$\hat{\theta}_n = \theta(\widehat{F}_n)\,, \tag{6.9}$$

where $\widehat{F}_n$ is the empirical distribution function computed from χ_n. Likewise we can define $\hat{\theta}_m^* = \theta(\widehat{F}_m^*)$, where $\widehat{F}_m^*$ is the empirical distribution function for χ_m^*. As we noted in Sect. 2, Bickel and Freedman (1981) showed that first-order properties of $\hat{\theta}_m^*$ are often robust against the value of m. In particular it is often the case that, for each $\varepsilon > 0$,

$$P(|\hat{\theta}_m^* - \hat{\theta}_n| > \varepsilon \,|\, \chi_n) \to 0\,, \quad P(|\hat{\theta}_n - \theta| > \varepsilon) \to 0 \tag{6.10}$$

as m and n diverge, where the first convergence is with probability 1. Compare (6.1). For example, (6.10) holds if θ is a moment, such as a mean or a variance, and if the sampling distribution has sufficiently many finite moments.

The definition (6.9) is conventionally used for a bootstrap estimator, and it does not necessarily involve simulation. For example, if $\theta = \int x\,dF(x)$ is a population mean then

$$\hat{\theta}_n = \int x\,d\widehat{F}_n(x) = \bar{X}\,, \quad \hat{\theta}_m^* = \int x\,d\widehat{F}_m^*(x) = \bar{X}^*$$

are the sample mean and resample mean, respectively. However, in a variety of other cases the most appropriate way of defining and computing $\hat{\theta}_n$ is in terms of the resample χ_n^*; that is, χ_m^* with $m = n$. Consider, for instance, the case where

$$\theta = P\left(X_{(n)} - X_{(n-1)} > X_{(n-1)} - X_{(n-2)}\right)\,, \tag{6.11}$$

in which, as in Sect. 6.1.2, we take $X_{(1)} < \ldots < X_{(n)}$ to be an ordering of the data in χ_n, assumed to have a common continuous distribution. For many sampling distributions, in particular distributions that lie in the domain of attraction of an extreme-value law, θ depends on n but converges to a strictly positive number as n increases.

In this example the bootstrap estimator, $\hat{\theta}_n$, of θ, based on a sample of size n, is defined by

$$\hat{\theta}_n = P\Big(X^*_{(n)} - X^*_{(n-1)} > X^*_{(n-1)} - X^*_{(n-2)} \,\Big|\, \chi_n\Big)\,, \tag{6.12}$$

where $X^*_{(1)} \leq \ldots \leq X^*_{(n)}$ are the ordered data in χ^*_n. Analogously, the bootstrap version, $\hat{\theta}^*_n$, of $\hat{\theta}_n$ is defined using the double bootstrap:

$$\hat{\theta}^*_n = P\Big(X^{**}_{(n)} - X^{**}_{(n-1)} > X^{**}_{(n-1)} - X^{**}_{(n-2)} \,\Big|\, \chi^*_n\Big)\,,$$

where $X^{**}_{(1)} \leq \ldots \leq X^{**}_{(n)}$ are the ordered data in $\chi^{**}_n = \{X^{**}_1, \ldots, X^{**}_n\}$, drawn by sampling randomly, with replacement, from χ^*_n. However, for the reasons given in the paragraph containing (6.5), property (6.10) fails in this example, no matter how we choose m. (The m in (6.2) is different from the m for the m-out-of-n bootstrap.) The bootstrap fails to model accurately the relationships among large order statistics, to such an extent that, in the example characterised by (6.11), $\hat{\theta}_n$ does not converge to θ.

This problem evaporates if, in defining $\hat{\theta}_n$ at (6.12), we take the resample χ^*_m to have size $m = m(n)$, where

$$m \to \infty \quad \text{and} \quad m/n \to 0 \tag{6.13}$$

as $n \to \infty$. That is, instead of (6.12) we define

$$\hat{\theta}_n = P\Big(X^*_{(m)} - X^*_{(m-1)} > X^*_{(m-1)} - X^*_{(m-2)} \,\Big|\, \chi_n\Big)\,, \tag{6.14}$$

where $X^*_1, \ldots, X^*_m$ are drawn by sampling randomly, with replacement, from χ_n. In this case, provided (6.5) holds, (6.2) is correct in a wide range of settings.

Deriving this result mathematically takes a little effort, but intuitively it is rather clear: By taking m to be of strictly smaller order than n we ensure that the probability that $X^*_{(m)}$ equals any given data value in χ_n, for example $X_{(n)}$, converges to zero, and so the difficulties raised in the paragraph containing (6.5) no longer apply. In particular, instead of (6.4) we have:

$$P(X^*_{(m-k)} = X_{(m-\ell)} \,|\, \chi_n) \to 0$$

in probability, for each fixed, nonnegative integer k and ℓ, as $n \to \infty$. Further thought along the same lines indicates that the conditional distribution of $X^*_{(m)} - X^*_{(m-1)}$ should now, under mild assumptions, be a consistent estimator of the distribution of $X_{(n)} - X_{(n-1)}$.

Bickel et al. (1997) gave a sequence of four counter-examples illustrating cases where the bootstrap fails, and provided two examples of the success of the bootstrap. The first two counter-examples relate to extrema, and so are closely allied to the example considered above. The next two treat, respectively, hypothesis testing and improperly centred U and V statistics, and estimating nonsmooth functionals of the population distribution function. Bickel et al. (1997) then developed a deep, general theory which allowed them to construct accurate and insightful approximations to bootstrap statistics $\hat{\theta}_n$, such as that at (6.9), not just in that case but also when $\hat{\theta}_n$ is defined using the m-out-of-n bootstrap, as at (6.14). This enabled them to show that, in a large class of problems for which (6.13) holds, the m-out-of-n bootstrap overcomes consistency problems inherent in the conventional n-out-of-n approach, and also to derive rates of convergence.

A reliable way of choosing m empirically is of course necessary if the m-out-of-n bootstrap is to be widely adopted. In many cases this is still an open problem, although important contributions were made recently by Bickel and Sakov (2008).

References

Bickel PJ, Freedman DA (1980) On Edgeworth expansions and the bootstrap. Unpublished manuscript

Bickel PJ, Freedman DA (1981) Some asymptotic theory for the bootstrap. Ann Stat 9:1196–1217

Bickel PJ, Freedman DA (1984) Asymptotic normality and the bootstrap in stratified sampling. Ann Stat 12:470–482

Bickel P, Sakov A (2008) On the choice of m in the m out of n bootstrap and confidence bounds for extrema. Stat Sin 18:967–985

Bickel PJ, Yahav JA (1988) Richardson extrapolation and the bootstrap. J Am Stat Assoc 83: 387–393

Bickel PJ, Götze F, van Zwet WR (1997) Resampling fewer than n observations: gains, losses, and remedies for losses. Stat Sin 7:1–31

Davison AC (1988) Discussion of papers by D.V. Hinkley and by T.J. DiCiccio and J.P. Romano. J R Stat Soc Ser B 50:356–357

Davison AC, Hinkley DV, Schechtman E (1986) Saddlepoint approximations in resampling methods. Biometrika 75:417–431

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26

Efron B (1988) Bootstrap confidence intervals: good or bad? (with discussion.) Psychol Bull 104:293–296

Efron B (1990) More efficient bootstrap computations. J Am Stat Assoc 85:79–89

Graham RL, Hinkley DV, John PWM, Shi S (1990) Balanced design of bootstrap simulations. J R Stat Soc Ser B 52:185–202

Gurney M (1963) The variance of the replication method for estimating variances for the CPS sample design. Memorandum, U.S. Bureau of the Census. Unpublished

Hall P (1978) Representations and limit theorems for extreme value distributions. J Appl Probab 15:639–644

Hall P (1986) On the number of bootstrap simulations required to construct a confidence interval. Ann Stat 14:1453–1462

Hall P (1990) Antithetic resampling for the bootstrap. Biometrika 76:713–724

Jeffreys Y, Jeffreys BS (1988) Methods of mathematical physics, 3rd edn. Cambridge University Press, Cambridge

Johns MV (1988) Importance sampling for bootstrap confidence intervals. J Am Stat Assoc 83:709–714

Jones HL (1956) Investigating the properties of a sample mean by employing random subsample means. J Am Stat Assoc 51:54–83

Mccarthy PJ (1966) Replication: an approach to the analysis of data from complex surveys. In: National Center for Health Statistics, Public Health Service (eds) Vital Health Statistics: Series 2. Public Health Service publication 1000, vol 14. U.S. Government Printing Office, Washington, DC

Mccarthy PJ (1969) Pseudo-replication: half samples. Rev Int Stat Inst 37:239–264

Shiue C-J (1960) Systematic sampling with multiple random starts. For Sci 6:42–50

Simon JC (1992) Barriers to adoption, and the future of resampling; resistances to using and teaching resampling. Unpublished

Singh K (1981) On the asymptotic accuracy of Efron's bootstrap. Ann Stat 9:1187–1195

Swanepoel JWH (1986) A note on proving that the (modified) bootstrap works. Commun Stat Ser A 15:3193–3203

The Annals of Statistics
1981, Vol. 9, No. 6, 1196–1217

SOME ASYMPTOTIC THEORY FOR THE BOOTSTRAP

BY PETER J. BICKEL[1] AND DAVID A. FREEDMAN[2]

University of California, Berkeley

Efron's "bootstrap" method of distribution approximation is shown to be asymptotically valid in a large number of situations, including t-statistics, the empirical and quantile processes, and von Mises functionals. Some counter-examples are also given, to show that the approximation does not always succeed.

1. Introduction. Efron (1979) discusses a "bootstrap" method for setting confidence intervals and estimating significance levels. This method consists of approximating the distribution of a function of the observations and the underlying distribution, such as a pivot, by what Efron calls the bootstrap distribution of this quantity. This distribution is obtained by replacing the unknown distribution by the empirical distribution of the data in the definition of the statistical function, and then resampling the data to obtain a Monte Carlo distribution for the resulting random variable. This method would probably be used in practice only when the distributions could not be estimated analytically. However, it is of some interest to check that the bootstrap approximation is valid in situations which are simple enough to handle analytically. Efron gives a series of examples in which this principle works, and establishes the validity of the approach for a general class of statistics when the sample space is finite.

In Section 2 of the present paper, it will be shown that the bootstrap works for means, and hence for pivotal quantities of the familiar "t-statistic" sort; an extension to multi-dimensional data will be made. Section 3 deals with U-statistics and other von Mises functionals, and suggests the wide scope of the theory. Section 4 deals with the empirical process: one application is setting confidence bounds for the theoretical distribution function, even if the latter has a discrete component. In Section 5, the quantile process will be bootstrapped, leading to confidence intervals for quantiles. Trimmed means and Winsorized variances are also studied. In Section 6 some examples will be given where the bootstrap fails, for instance, when estimating θ from variables uniformly distributed over $[0, \theta]$.

Some of the problems discussed in this paper have been studied independently by Singh (1981).

2. Bootstrapping the mean. Let $X_1, X_2, \cdots, X_n$ be independent random variables with common distribution function F. Assume that F has finite mean μ and variance σ^2, both unknown. The conventional estimate for μ is the sample average, denoted here by μ_n. To analyze the sampling error in μ_n, it is customary to compute the sample standard deviation s_n, defined as

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu_n)^2.$$

Received July 14, 1980; revised March 11, 1981.

[1] Research partially supported by Office of Naval Research, Contract N-00014-80-C-0163, and the Adolph and Mary Sprague Miller Foundation.

[2] Research partially supported by National Science Foundation Grant MCS-80-02535.

AMS 1970 *subject classifications.* Primary 62E20; secondary 62G05, 62G15.

Key words and phrases. Bootstrap, resampling, asymptotic theory.

By the Classical Central Limit Theorem, the distribution of the pivotal quantity

$$Q_n = \sqrt{n}(\mu_n - \mu)/s_n$$

tends weakly to $N(0, 1)$. So, in this situation, the asymptotics are known. However, there is some theoretical interest in seeing how the bootstrap would perform.

Let F_n be the empirical distribution of $X_1, \cdots, X_n$, putting mass $1/n$ on each X_i. The next step in the bootstrap method is to resample the data. Given $(X_1, \cdots, X_n)$, let $X_1^*, \cdots, X_m^*$ be conditionally independent, with common distribution F_n. We have allowed the resample size m to differ from the number n of data points, to estimate the distribution of the bootstrap pivotal quantity $Q_m^* = \sqrt{m}(\mu_m^* - \mu_n)/s_m^*$, where $\mu_m^* = (1/m) \sum_{i=1}^m X_i^*$ and $s_m^* = (1/m) \sum_{i=1}^m (X_i^* - \mu_m^*)^2$.

In the resampling, the n data points $X_1, \cdots, X_n$ are treated as a population, with distribution function F_n and mean μ_n; and μ_m^* is considered as an estimator of μ_n. First, take $m = n$. The idea is that the behavior of the bootstrap pivotal quantity Q_n^* mimics that of Q_n. Thus, the distribution of Q_n^* could be computed from the data and used to approximate the unknown sampling distribution of Q_n. Or even more directly, the bootstrap distribution of $\sqrt{n}(\mu_n^* - \mu_n)$ could be used to approximate the sampling distribution of $\sqrt{n}(\mu_n - \mu)$. Either approach would lead to confidence intervals for μ, and would be useful if the Central Limit Theorem were not available, and if the bootstrap approximation were valid.

Now take $m \neq n$. The resample size m does have some statistical import. For instance, a sample of size n can be bootstrapped to see what would happen with a sample of size n^2, or $\sqrt{n}$, or 10. Furthermore, with m and n free to vary separately, the second-moment condition in Theorem 2.1 becomes quite natural. If m goes to infinity first, then the conditional law of $\sqrt{m}(\mu_m^* - \mu_n)$ tends to normal, with mean 0 and variance s_n^2. As n tends to infinity, this converges if and only if s_n^2 does.

Mathematically, there is something rather delicate even about the present simple case, with $m = n$. Comparing the classical $\sqrt{n}(\mu_n - \mu)$ with the bootstrap $\sqrt{n}(\mu_n^* - \mu_n)$, the parameter μ is replaced by μ_n. But this change is of the critical order of magnitude, namely $1/\sqrt{n}$, and cannot be ignored. However, there is a second error: the X's have been replaced by X^*'s. In fact, these two errors cancel each other to a large extent. Our proof will make this idea precise, by showing that the distribution of the pivot does not change much if the empirical F_n is replaced by the theoretical F. The theorem is an asymptotic one, so the data $X_1, \cdots, X_n$ should be visualized as the beginning segment of an infinite series.

THEOREM 2.1. *Suppose $X_1, X_2, \cdots$ are independent, identically distributed, and have finite positive variance σ^2. Along almost all sample sequences $X_1, X_2, \cdots$, given $(X_1, \cdots, X_n)$, as n and m tend to ∞:*

(*a*) *The conditional distribution of $\sqrt{m}(\mu_m^* - \mu_n)$ converges weakly to $N(0, \sigma^2)$.*

(*b*) *$s_m^* \to \sigma$ in conditional probability: that is, for ϵ positive,*

$$P\{|s_m^* - \sigma| > \epsilon | X_1, \cdots, X_n\} \to 0 \text{ a.s.}$$

Relations (a) and (b) imply that the asymptotic distribution of the bootstrap pivot Q_n^* coincides with the classical one: convergence to the standard normal holds. There are several equivalent ways to prove these results. We choose an argument which is qualitative, but requires some machinery. Let Γ_2 be the set of distribution functions G satisfying $\int x^2\, dG(x) < \infty$, and introduce the following notion of convergence in Γ_2:

$$G_\alpha \Rightarrow G \quad \text{iff} \quad G_\alpha \to G \text{ weakly} \quad \text{and} \quad \int x^2\, dG_\alpha(x) \to \int x^2\, dG(x).$$

The strong law implies

$$\text{(2.1)} \qquad F_n \Rightarrow F \quad \text{along almost all sample sequences.}$$

The conclusions of the theorem hold along any such sample sequence.

Our notion of convergence in Γ_2 is metrizable, for instance, by a "Mallows metric" d_2. The d_2-distance between G and H in Γ_2 is defined as follows: $d_2(G, H)^2$ is the infimum of $E\{(X - Y)^2\}$ over all joint distributions for the pair of random variables X and Y whose fixed marginal distributions are G and H respectively. This metric was introduced in Mallows (1972) and Tanaka (1973); it is related to the Vassershtein metrics of Dobrushin (1970), Major (1978), or Vallender (1973). For a detailed discussion of d_2, see Section 8 of the present paper.

Now let $Z_1(G), \cdots, Z_m(G)$ be independent random variables, with common distribution function G. Let $G^{(m)}$ be the distribution of

$$S_m(G) = m^{-1/2} \sum_{j=1}^m [Z_j(G) - E\{Z_j(G)\}].$$

If $G \in \Gamma_2$, so is $G^{(m)}$. By Lemma 3 of Mallows (1972),

$$d_2[G^{(m)}, H^{(m)}] \leqq d_2[G, H]. \tag{2.2}$$

Also see Lemma 8.7 below, and (8.2).

Proof of Theorem 2.1, *Part a.* The bootstrap construction can be put into present notation as follows: conditionally, the law of $\sqrt{m}(\mu_m^* - \mu_n)$ is just $F_n^{(m)}$. But F_n is close to F in the d_2-metric on Γ_2, by (2.1). So $F_n^{(m)}$ is close to $F^{(m)}$ by (2.2). Now use the ordinary Central Limit Theorem on $F^{(m)}$.

Part b. This can be proved the same way. Let Γ_1 be the set of G's with $\int |x| G(dx) < \infty$, and define the metric d_1 on Γ_1 as the infimum of $E\{|X - Y|\}$ over all pairs of random variables X and Y with marginal distributions F and G respectively. Let $G^{(m)}$ be the distribution of $(1/m) \sum_{j=1}^m Z_j(G)$. The requisite analog of (2.2) is

$$d_1[G^{(m)}, H^{(m)}] \leqq d_1[G, H]. \tag{2.3}$$

For details on d_1, See Section 8, especially Lemma 8.6. □

The following generalization to higher dimensions may be of some interest. Let $\|\cdot\|$ denote length in R^k.

Theorem 2.2. *Let $X_1, X_2, \cdots$ be independent, with common distribution in R^k. Suppose $E\{\|X_1\|^2\} < \infty$. Let F_n be the empirical distribution of $X_1, \cdots, X_n$. Given $X_1, \cdots, X_n$, let $X_1^*, \cdots, X_m^*$ be conditionally independent, with common distribution F_n. Along almost all sample sequences, as m and n tend to infinity:*

(*a*) *The conditional distribution of*

$$\sqrt{m}\left(\frac{1}{m} \sum_{j=1}^m X_j^* - \frac{1}{n} \sum_{i=1}^n X_i\right)$$

converges weakly to the k-dimensional normal distribution with mean 0, *and variance-covariance matrix equal to the theoretical variance-covariance matrix of X_1.*

(*b*) *The empirical variance-covariance matrix of $X_1^*, \cdots, X_m^*$ converges in conditional probability to the theoretical variance-covariance matrix of X_1.*

The requisite metrics are developed in Section 8. If, e.g., $E\{\|X_1\|^4\} < \infty$ then the estimated variance-covariance matrix can be bootstrapped in turn, and so on. We do not pursue this further.

Efron considers the possibility of resampling not from F_n, but from some other estimator, call it $\tilde{F}_n$, of F. The argument for Theorem 2.1 shows that this works too, provided $\tilde{F}_n \Rightarrow F$ in Γ_2, i.e., $\tilde{F}_n$ gets F almost right in the weak topology, and also gets the second moment almost right.

As a lead-in to the treatment of U-statistics in Section 3, fix a function h on $(-\infty, \infty)$ and let Γ_h be the set of distribution functions G satisfying

$$\int h^2(x)\, dG(x) < \infty.$$

Then the estimator $(1/n) \sum_{i=1}^n h(X_i)$ can be bootstrapped, provided the distribution of the X's is in Γ_h. The relevant notion of convergence seems to be this:

$$G_\alpha \Rightarrow G \text{ in } \Gamma_h \quad \text{iff} \quad \int h^2\, dG_\alpha \to \int h^2\, dG, \quad \text{and} \quad \int \theta(h)\, dG_\alpha \to \int \theta(h)\, dG$$

for all bounded continuous functions θ on the line. This just repeats the theorem, in a form more convenient for use in Section 3.

Let $\tilde{F}_n$ be an estimator of F. We continue to assume that $F \in \Gamma_h$. Consider bootstrapping $(1/n) \sum_{i=1}^n h(X_i)$, but resampling from $\tilde{F}_n$ rather than F_n. When will this be asymptotically right? What is needed is the analog of the strong law of large numbers,

$$\int v(x)\, d\tilde{F}_n(x) \to \int v(x)\, dF(x) \quad \text{a.s.} \tag{2.4}$$

whenever $\int |v(x)|\, dF(x) < \infty$. The exceptional null set may depend on v. In particular, suppose $\tilde{F}_n = F_{\hat{\theta}_n}$ where F_θ is some parametric model under consideration and $\hat{\theta}_n(X_1, \cdots, X_n)$ is an estimate of θ. Efron calls this the parametric bootstrap. Then (2.4) holds when $F = F_{\theta_0}$ if $\hat{\theta}_n$ is strongly consistent and the map $\theta \to \int v(x)\, dF_\theta(x)$ is continuous at θ_0 whenever $\int |v(x)|\, dF_{\theta_0}(x) < \infty$.

To close this section, we set our results in the general context introduced by Efron. He considers real valued functions $Z_n(\cdot, \cdot)$ on $Z^n \times \mathscr{F}$ where $\mathscr{F}$ is a set of probability distributions on R containing the "true" F and all distributions with finite support. The bootstrap works if the conditional distribution of $Z_n\{(X_1^*, \cdots, X_n^*), F_n\}$ is close to the distribution of $Z_n\{(X_1, \cdots, X_n), F\}$. We interpret this as follows: If the law of $Z_n\{(X_1, \cdots, X_n), F\}$ tends weakly to a limit as $n \to \infty$, then the conditional distribution of $Z_m \cdot \{(X_1^*, \cdots, X_m^*), F_n\}$ given $(X_1, \cdots, X_n)$ tends weakly to the same limit law with probability one as $m, n \to \infty$. Theorem 2.1 shows this for

$$Z_n\{(X_1, \cdots, X_n), F\} = n^{1/2}\left\{n^{-1} \sum_{i=1}^n X_i - \int x\, dF(x)\right\}.$$

The present notion of convergence is stronger than Efron's, who requires only that the conditional distributions converge weakly to the same limit law in probability. Efron has established convergence in his sense for the mean, when F has finite support.

3. Bootstrapping von Mises functionals. Suppose $X_1, \cdots, X_n$ are independent identically distributed p vectors. Many pivots of interest which have limiting normal distributions can be written in the form

$$\frac{n^{1/2}\{g(S_n/n) - g(\mu)\}}{v(T_n/n)}$$

where $g : R^k \to R$, $v : R^\ell \to R$,

$$S_n = \textstyle\sum_{i=1}^n h(X_i), \tag{3.1}$$

$$T_n = \textstyle\sum_{i=1}^n r(X_i), \tag{3.2}$$

$h : R^p \to R^k$, $r : R^p \to R^\ell$, and

$$\mu = Eh(X_1), \qquad \nu = Er(X_1). \tag{3.3}$$

The asymptotic theory for such things is, of course, based on linearization for the numerator

$$n^{1/2}\left\{g\left(\frac{S_n}{n}\right) - g(\mu)\right\} = \dot{g}(\mu)n^{1/2}\left(\frac{S_n}{n} - \mu\right)^T + o_p(1) \tag{3.4}$$

provided that $E \| h(X_1) \|^2 < \infty$, g has a total differential $\dot{g}_{1\times k}$ at μ, and for the denominator that v is continuous at ν in the sense

$$v\left(\frac{T_n}{n}\right) = v(\nu) + o_p(1). \tag{3.5}$$

The bootstrap commutes with smooth functions in exactly the same way. Let

$$\tilde{S}_n = \textstyle\sum_{i=1}^n h(Y_i^*), \qquad \tilde{T}_n = \textstyle\sum_{i=1}^n r(Y_i^*).$$

If $E \| h(X_1) \|^2 < \infty$ and $\dot{g}$ exists in a neighborhood of μ and is continuous at μ then,

$$n^{1/2}\left\{g\left(\frac{\tilde{S}_n}{n}\right) - g\left(\frac{S_n}{n}\right)\right\} = \dot{g}(\mu)n^{1/2}\left(\frac{S_n}{n} - \frac{\tilde{S}_n}{n}\right)^T + \Delta_n \tag{3.6}$$

where $\Delta_n \to 0$ in conditional probability and, of course, if v is continuous

$$v\left(\frac{\tilde{T}_n}{n}\right) \to v(\nu) \tag{3.7}$$

in conditional probability. The proof of (3.6) in a more general setting is given in Lemma 8.10 below.

Suppose now that g is a functional $g : \mathscr{F} \to R$ where $\mathscr{F}$ is a convex set of probability measures on R^m including all point masses and F. Suppose also that g is Gâteaux differentiable at F with derivative $\dot{g}(F)$ representable as an integral

$$\dot{g}(F)(G - F) = \frac{\partial}{\partial\epsilon} g(F + \epsilon(G - F))\big|_{\epsilon=0} = \int \psi(x, F)\, dG(x) \tag{3.8}$$

where necessarily

$$\int \psi(x, F)\, dF(x) = 0. \tag{3.9}$$

Such g are often called von Mises functionals. Asymptotic normality results in nonparametric statistics relate to quantities of the form $n^{1/2}\{g(F_n) - g(F)\}$ or asymptotically equivalent quantities. The result we usually want and get is that $n^{1/2}\{g(F_n) - g(F)\}$ and $n^{1/2} \int \psi(x, F)\, d(F_n - F)$ have the same $N(0, \int \psi^2(x, F)\, dF)$ limit law. As Reeds (1976) indicates, this reflects a general Taylor approximation

$$g(F_n) - g(F) = \dot{g}_F(F_n - F) + \Delta_n(F_n, F) \tag{3.10}$$

where

$$\Delta_n(F_n, F) = o_p(g_F(F_n - F)).$$

It is natural to hope that if we let G_n be the empirical d.f. of $X_1^*, \cdots, X_n^*$, then

$$g(G_n) - g(F_n) = \dot{g}_{F_n}(G_n - F_n) + \Delta_n(G_n, F_n),$$

where for almost all $X_1, X_2, \cdots$

$$n^{1/2}\Delta_n(G_n, F_n) \to 0 \tag{3.11}$$

in conditional probability, and thence that the conditional law of

$$n^{1/2}\dot{g}_{F_n}(G_n - F_n) = n^{-1/2} \textstyle\sum_{i=1}^n \psi(X_i^*, F_n) \text{ tends to } N\left(0, \displaystyle\int \psi^2(x, F)\, dF(x)\right). \tag{3.12}$$

Simple conditions for the validity of (3.11) can be formulated using the theory of compact differentiation as in Reeds (1976). However, verification of these conditions in particular situations poses the same requirements for special arguments as in Reeds' verification of various examples of (3.10). Moreover, whereas convergence in law under F of $\int \psi(x, F)\, dF_n$ is immediate if $\int \psi^2(x, F)\, dF < \infty$, further continuity conditions on ψ as a function of F seem necessary to ensure that the conditional distributions of $\int \psi(x, F_n)\, dG_n$ tend weakly to $N(0, \int \psi^2(x, F)\, dF(x))$.

The simplest conditions sufficient to guarantee this behavior seem to be

$$\text{i)} \int \psi^2(x, F)\, dF(x) < \infty.$$

$$\text{ii)} \int (\psi(x, F_n) - \psi(x, F))^2\, dF_n \to 0 \text{ a.s.}$$

Condition (ii) implies that for almost all $X_1, X_2, \cdots,$

$$n^{-1/2} \sum_{i=1}^n \left[\psi(X_i^*, F_n) - \left\{ \psi(X_i^*, F) - \int \psi(x, F)\, dF_n \right\} \right] \to 0$$

in conditional probability, while condition (i) ensures the satisfactory behavior of $n^{-1/2}\Sigma\psi(X_i^*, F) - \int \psi(x, F)\, dF_n$. These conditions are exploited in Theorem 3.1 below.

We pursue these general considerations slightly in Section 8. Here we content ourselves with checking the bootstrap for the simplest nonlinear von Mises functionals

$$g(H) = \int\int \omega(x, y)\, dH(x)\, dH(y) \tag{3.13}$$

where $\omega(x, y) = \omega(y, x)$ and H is such that $g(H)$ is well defined. In particular,

$$g(F_n) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \omega(X_i, X_j).$$

A closely related statistic of interest is the U-statistic of order 2 defined by

$$g_n(F_n) = \binom{n}{2}^{-1} \sum_{i<j} \omega(X_i, X_j) = \frac{n}{n-1} g(F_n) - \frac{1}{n(n-1)} \sum_{i=1}^n \omega(X_i, X_i). \tag{3.14}$$

It is well known (von Mises, 1947) that if

$$\int \omega^2(x, y)\, dF(x)\, dF(y) < \infty \tag{3.15}$$

and

$$\int \omega^2(x, x)\, dF(x) < \infty, \tag{3.16}$$

then

$$n^{1/2}\{g(F_n) - g(F)\} \text{ tends weakly to } N(0, \sigma^2) \tag{3.17}$$

where

$$\sigma^2 = 4\left[\int \left\{ \int \omega(x, y)\, dF(y) \right\}^2 dF(x) - g^2(F) \right]. \tag{3.18}$$

This is in accord with (3.8) and (3.10), since in this case

$$\psi(x, F) = 2\left\{ \int \omega(x, y)\, dF(y) - g(F) \right\}. \tag{3.19}$$

THEOREM 3.1 *If* (3.15) *and* (3.16) *hold, and g is given by* (3.13) *and* σ^2 *by* (3.18), *then for almost all* $X_1, X_2, \cdots,$ *given* $(X_1, \cdots, X_n)$,

$$n^{1/2}\{g(G_n) - g(F_n)\} \quad \text{converges weakly to} \quad N(0, \sigma^2).$$

PROOF. Define ψ and Δ_n as in (3.19) and (3.10). Then we will establish that (3.11) and (3.12) hold.

PROOF OF CLAIM (3.11). $\Delta_n(G_n, F_n) = \int\int \omega(x, y)\, d(G_n - F_n)(x)\, d(G_n - F_n)(y)$. By an inequality of von Mises (1947) (see also Hoeffding, 1948),

$$E\{\Delta_n^2(G_n, F_n)|X_1, \cdots, X_n\} \leqq n^{-2}\left\{C_1 \int\int \omega^2(x, y)\, dF_n\, dF_n + \frac{C_2}{n}\int \omega^2(x, x)\, dF_n\right\}.$$

where C_1 and C_2 are universal constants. Now

$$\int \omega^2(x, x)\, dF_n \to E\omega^2(X_1, X_1)$$

$$\int\int \omega^2(x, y)\, dF_n\, dF_n = \left(\frac{n}{n-1}\right)^2 \binom{n}{2}^{-1} \sum_{i<j} \omega^2(X_i, X_j)$$

$$+ n^{-2}\sum_i \omega^2(X_i, X_i) \to E\omega^2(X_1, X_2)$$

almost surely by the strong law of large numbers, as generalized to U-statistics (see Berk, 1966, page 56) and (3.11) follows.

PROOF OF CLAIM (3.12). As we noted earlier, it is enough to show that

$$\int \{\psi(x, F_n) - \psi(x, F)\}^2\, dF_n \to 0$$

with probability 1. But,

$$\int \{\psi(x, F_n) - \psi(x, F)\}^2\, dF_n(x) = n^{-1}\sum_i \{\psi(X_i, F_n) - \psi(X_i, F)\}^2$$

$$= n^{-1}\sum_i \left\{n^{-1}\sum_j \omega(X_i, X_j) - \int \omega(X_i, y)\, dF(y)\right\}^2$$

$$= n^{-3}\sum_{i,j,k} \omega(X_i, X_j)\omega(X_i, X_k)$$

$$- 2n^{-2}\sum_{i,j} \omega(X_i, X_j)\int \omega(X_i, y)\, dF$$

$$+ n^{-1}\sum_i \left\{\int \omega(X_i, y)\, dF\right\}^2.$$

By an argument using a strong law of large numbers for U-statistics, these last three terms tend with probability 1 to

$$E\omega(X_1, X_2)\omega(X_1, X_3),\ -2E[\omega(X_1, X_2)E\{\omega(X_1, X_2)|X_2\}], \text{and } E[E^2\{\omega(X_1, X_2)|X_2\}],$$

respectively. The sum of these numbers is 0 and claim (3.12) and the theorem follow. □

If $E\omega^2(X_1, X_2) < \infty$ and $E\omega^2(X_1, X_1) < \infty$, the conclusion of Theorem 3.1 clearly holds for the bootstrap distribution of the U-statistic $g_n(F_n)$ and, more generally, any convex combination of $g_n(F_n)$ and $n^{-1}\sum \omega(X_i, X_i)$ where the weight on $g_n(F_n)$ tends to 1. Failure of the conditions, however, can cause failure of the bootstrap (see Section 6).

As an example of the applicability of this result, it is valid to bootstrap the distribution of Wilcoxon's one sample statistic

$$\left\{\frac{n^{1/2}(n+1)}{2}\right\}^{-1} \sum_{i\leq j} \{I(X_i + X_j > 0) - P(X_i + X_j > 0)\}$$

in order, for instance, to obtain approximations to its power.

Extensions of the theorem to the von Mises statistics corresponding to U-statistics of arbitrary order, vector U-statistics, U-statistics based on several samples, etc., is straightforward, provided, however, that the hypotheses appropriate to the von Mises statistics, as in Fillipova (1962), are kept.

Extending a remark made in Section 2, we can bootstrap U-statistics by resampling from a general $\{\tilde{F}_n\}$, provided that $\{\tilde{F}_n\}$ possesses a property analogous to the strong law of large numbers for U-statistics, viz.,

$$\int \cdots \int v(x_1, \cdots, x_k)\, dF_n(x_1) \ldots dF_n(x_k) \to \int \cdots \int v(x_1, \cdots, x_k)\, dF(x_1) \ldots dF(x_k) \text{ a.s.}$$

if
$$\int |v(x_1, \cdots, x_k)|\, dF(x_1) \ldots dF(x_k) < \infty.$$

4. Bootstrapping the empirical process. The object of this section is to bootstrap the empirical process, (Theorem 4.1), and to obtain a fixed-width confidence band for the population distribution function which is valid even when the latter has a discrete component (Corollary 4.2). We first give two preliminary lemmas and then recall notions of weak convergence. Throughout this section, B is a Brownian bridge on [0, 1]. Theorem 3 of Komlos, Major and Tusnady (1975) implies the following result.

LEMMA 4.1 *There exist, on a sufficiently rich probability space, independent random variables $U_1, U_2, \ldots$ with common distribution uniform on* [0, 1], *and a Brownian bridge B on* [0, 1] *with the following property. Let H_m be the empirical distribution function of* $U_1, \cdots, U_m$ and let

$$B_m(u) = m^{1/2}\{H_m(u) - u\} \quad \text{for } 0 \leqq u \leq 1.$$

Then for some constant K_1, and $\epsilon_m = (\log m)/m^{1/2}$

$$P\{\|B_m - B\| \geqq K_1\epsilon_m\} \leqq K_1\epsilon_m.$$

To state the next result, which is an integrated form of Levy's modulus of continuity, let

$$\omega(\delta, f) = \sup\{|f(s) - f(t)| : |t - s| \leqq \delta\} \tag{4.1}$$

$$\begin{aligned} h(\delta) &= \left(\delta \log \frac{1}{\delta}\right)^{1/2} && \text{for } 0 \leqq \delta \leqq ½ \\ &= h(1/2) && \text{for } \delta \geqq ½ \end{aligned} \tag{4.2}$$

LEMMA 4.2 *There is a constant K_2 such that $E\{\omega(\delta, B)\} \leqq K_2 h(\delta)$ for $0 < \delta \leqq ½$.*

PROOF. Represent B as

$$B(u) = W(u) - uW(1) \quad \text{for} \quad 0 \leqq u \leqq 1,$$

where W is a Wiener process on $[0, \infty)$. Now

$$\omega(\delta, B) \leqq \omega(\delta, W) + \delta |W(1)|.$$

So it is enough to prove the lemma with W in place of B. Abbreviate

$$M_{k\delta} = \sup_s \{|W(s) - W(k\delta)| : k\delta \leqq s \leqq (k+1)\delta\}.$$

Let K be the integer part of $1/\delta$. By the triangle inequality,

$$\omega(\delta, W) \leqq 3 \max_k \{M_{k\delta} : 0 \leqq k \leqq K\}.$$

Of course, the $M_{k\delta}$ are independent and identically distributed, so

$$E\{\omega(\delta, W)\} = \int_0^\infty P\{\omega(\delta, W) > x\}\, dx \leqq 3 \int_0^\infty [1 - \{1 - P(M_{o\delta} > x)\}^{K+1}]\, dx.$$

If $x < 2^{1/2}h(\delta)$, the integrand may be replaced by the trivial upper bound of 1. The integral over bigger x's is negligible for small δ; this may be seen by estimating the integrand as follows:

$$1 - (1-p)^{K+1} \leqq (K+1)p \qquad \text{for } 0 \leqq p \leqq 1$$

$$P\{M_{o\delta} > x\} \leqq 4(\delta/2\pi)^{1/2}x^{-1}e^{-x^2/2\delta}$$

and then making the change of variables $y = \delta^{-1/2}x$. □

Let D be the space of all real-valued functions f on $[-\infty, \infty]$, such that f vanishes continuously at $\pm\infty$, and is right continuous with left limits on $(-\infty, \infty)$. Give D the Skorokhod topology. Let Γ be the set of all distribution functions, in the sup norm. For $G \in \Gamma$, let $Z_1(G), \cdots, Z_m(G)$ be independent with common distribution G. Let G_m be the empirical distribution of $Z_1(G), \cdots, Z_m(G)$, and set

$$W_{Gm}(t) = \sqrt{m}[G_m(t) - G(t)] \qquad \text{for } -\infty < t < \infty, \tag{4.3}$$

extended to vanish at $\pm\infty$. Let $\psi_m(G)$ be the distribution of the process W_{Gm}. Thus, $\psi_m(G)$ is a probability measure on D. In this notation, the usual invariance principle states that $\psi_m(G)$ tends weakly to the law of $B(G)$ as $m \to \infty$, where B is the Brownian bridge, and $B(G)(t, \omega) = B\{G(t), \omega\}$.

The weak topology on the space of probability measures on D is metrized by a dual Lipschitz metric as follows. Let γ metrize the Skorokhod topology on D, and in addition satisfy

$$\gamma(f, g) \leqq \|f - g\| \wedge 1. \tag{4.4}$$

Here f and g are elements of D, i.e., function on $[-\infty, \infty]$, and $\|\cdot\|$ is the sup norm. Now

$$\rho(\pi, \pi') = \sup_\theta \left| \int_D \theta\tau d\pi - \int_D \theta\tau d\pi' \right| \tag{4.5}$$

where π and π' are probability measures on D, and θ runs through the functions on D which are uniformly bounded by 1 and satisfy the Lipschitz condition

$$|\theta(f) - \theta(g)| \leqq \gamma(f, g).$$

PROPOSITION 4.1. *There exists a universal constant C such that*

$$\rho[\psi_m(F), \psi_m(G)] \leqq C[\epsilon_m + h(\|F - G\|)],$$

where $\epsilon_m = m^{-1/2}\log m$ *and h was defined in* (4.2).

PROOF. Recall B_m from Lemma 4.1 Clearly, $\psi_m(F)$ and $\psi_m(G)$ are the probability distributions induced on D by $B_m(F)$ and $B_m(G)$ respectively. By the definition (4.5) of the dual Lipschitz metric ρ,

$$\rho[\psi_m(F), \psi_m(G)] \leqq \sup_\theta E\{|\theta[B_m(F)] - \theta[B_m(G)]|\} \leqq E\{\gamma[B_m(F), B_m(G)]\}.$$

Now (4.4) implies

$$E\{\gamma[B_m(F), B_m(G)]\} \leqq E\{\|B_m(F) - B_m(G)\| \wedge 1\} \tag{4.6}$$

Since $\|f - g\| \wedge 1$ is a metric, the triangle inequality implies

$$E\{\gamma[B_m(F), B_m(G])\} \leqq 2E\{\|B_m - B\| \wedge 1\} + E\{\omega(\|F - G\|, B)\}. \tag{4.7}$$

Now use Lemma 4.1 to estimate the first term on the right in (4.7):

$$E\{\|B_m - B\| \wedge 1\} \leqq K_1\epsilon_m + P\{\|B_m - B\| > K_1\epsilon_m\} \leqq 2K_1\epsilon_m.$$

The second term on the right in (4.7) can be estimated by Lemma 4.2. □

Return now to the setting of Section 2, but with no moment condition. There is a sample of size n from an unknown distribution function F, which is to be estimated by the empirical distribution function F_n. Given $X_1, \cdots, X_n$, let $X_1^*, \cdots, X_m^*$ be conditionally independent, with common distribution F_n. Let F_{nm} be the empirical distribution function of $X_1^*, \cdots, X_m^*$. And let

$$W_{nm}(t) = \sqrt{m}\{F_{nm}(t) - F_n(t)\} \qquad \text{for } -\infty < t < \infty, \tag{4.8}$$

extended to vanish at $\pm\infty$. The next result is the bootstrap analog of the invariance principle, which states that $\sqrt{n}\,(F_n - F)$ converges weakly to $B(F)$ as $n \to \infty$. No conditions are imposed on F; as usual, B is the Brownian bridge on [0, 1].

THEOREM 4.1. *Along almost all sample sequences, given* $(X_1, \cdots, X_n)$, *as n and m tend to infinity,* W_{nm} *converges weakly to* $B(F)$.

PROOF. This is almost immediate from Proposition 4.1. Conditionally, $W_{nm} = W_{F_n m}$ has the law $\psi_m(F_n)$, and $\|F_n - F\| \to 0$ a.s. by the Glivenko-Cantelli lemma, so $\psi_m(F_n)$, is nearly $\psi_m(F)$. The latter is almost the law of $B(F)$ by the ordinary invariance principle. Indeed, the argument shows that the ρ-distance between $\psi_m(F_n)$ and the law of $B(F)$ is at most a universal constant times $\epsilon_m + h(\|F_n - F\|)$. □

COROLLARY 4.1. *For almost all* $X_1, X_2, \cdots$, *given* $(X_1, \cdots, X_n)$, *as n and m tend to infinity,* $\|F_{nm} - F\|$ *tends to* 0 *in probability. Here,* F_{nm} *is the empirical distribution of the resampled data, as defined above.*

We now consider confidence bands for F which will be valid even when F has a discrete component.

COROLLARY 4.2. *Suppose F is nondegenerate. Fix* α *with* $0 < \alpha < 1$. *Choose* $c(F_n)$ *from the bootstrap distribution so that*

$$P\{n^{1/2} \sup_x |F_{nn}(x) - F_n(x)| \leqq c_n(F_n) | X_1, \cdots, X_n\} \to 1 - \alpha.$$

Then

$$P\{n^{1/2} \sup_x | F_n(x) - F(x) | \leqq c_n(F_n)\} \to 1 - \alpha.$$

PROOF. Indeed, $c_n(F_n)$ must converge to the $(1-\alpha)$-point of the law of $\sup_x | B(F(x))|$, which is continuous: see Lemma 8.11 below. So, $F_n \pm c_n(F_n)$ is the desired band.

Preliminary calculations suggest that the mapping $F \to \psi_m(F)$ is uniformly equicontinuous, in the sense that there is a function $q(t) \to 0$ as $t \to 0$, and for all m, F and G:

$$\rho[\psi_m(F), \psi_m(G)] \leqq q(\|F - G\|).$$

The argument rests on the following inequality, which may be of independent interest. Suppose F and G concentrate on [0, 1] and $\|F - G\| < \delta$. Then

$$\text{Lebesgue measure of } \{t: 0 \leqq t \leqq 1 \quad \text{and} \quad |F^{-1}(t) - G^{-1}(t)| > \sqrt{\delta}\} < \sqrt{\delta}.$$

This is immediate from Chebychev's inequality; see (8.1).

Suppose the resampling is from another estimator $\tilde{F}_n$ for F. Bootstrapping may still be valid. Given $(X_1, \cdots, X_n)$, it can be shown that $W_{\tilde{F}_n m}$ tends weakly to $B(F)$ as m and n tend to ∞, provided $\tilde{F}_n \to F$ a.s. in the sup norm. Here $W_{\tilde{F}_n m}$ was defined in (4.3). This result can even be proven under the weaker hypothesis, that $\tilde{F}_n \to F$ a.s. in the Skorokhod topology.

5. The quantile process. Another interesting process in terms of which various statistics and pivots can be defined naturally is the quantile process Q_n which we define on (0, 1) by

$$Q_n(t) = n^{1/2}\{F_n^{-1}(t) - F^{-1}(t)\}$$

where the inverse of a distribution function H is given, in general, by

$$H^{-1}(t) = \inf\{x{:}H(x) \geqq t\}.$$

Our aim in this section is to justify the bootstrapping of this process. Applications which will be sketched briefly after the theorem include confidence intervals for the median and pivots based on trimmed means and Winsorized variances.

For convenience, throughout this section we use $\circ$ to denote composition. For example, $f \circ F^{-1}$ means $f(F^{-1})$.

It is well known (see Bickel, 1966, for example) that given $0 < t_0 \leqq t_1 < 1$, if

(5.1) F has continuous positive density f on R,

then

(5.2) Q_n tends weakly to $B/f \circ F^{-1}$ in the space of probability measures on $D[t_0, t_1]$.

Write G_n for F_{nn} as defined for (4.8) and let

$$Q_n = n^{1/2}(G_n^{-1} - F_n^{-1}).$$

THEOREM 5.1. *If* (5.1) *holds, then along almost all sample sequences* $X_1, X_2, \cdots$, *given* $(X_1, \cdots, X_n)$, Q_n *converges weakly to* $B/(f \circ F^{-1})$ *in the sense of weak convergence for probability measures on* $D[t_o, t_1]$.

PROOF. An equicontinuity argument does not work here since the behavior of the quantile process depends on the density of the limit distribution. This is also the reason we take $m = n$. We present a relatively ad hoc modification of an argument due to Pyke and Shorack (1968).

It is convenient to denote the sup norm in $D[t_o, t_1]$ by $\|\cdot\|$. Write

$$Q_n = n^{1/2}\frac{(F\circ G_n^{-1} - F\circ F_n^{-1})}{R_n},$$

where

$$R_n = \frac{F\circ G_n^{-1} - F\circ F_n^{-1}}{G_n^{-1} - F_n^{-1}}.$$

Continue by writing

$$\begin{aligned} n^{1/2}(F\circ G_n^{-1} - F\circ F_n^{-1}) &= n^{1/2}[\{(F_n\circ G_n^{-1} - F\circ G_n^{-1}) - (F_n\circ F_n^{-1} - F\circ F_n^{-1})\} \\ &\quad + \{G_n\circ G_n^{-1} - F_n\circ G_n^{-1}\}] \\ &\quad - n^{1/2}(F_n\circ F_n^{-1} - G_n\circ G_n^{-1}). \end{aligned} \tag{5.3}$$

Let the probability space be rich enough to support the processes B_n and B of Lemma 4.1 as well as another pair $(\tilde{B}_n, \tilde{B})$ with the same distribution as (B_n, B) and independent of them.

We now represent $n^{1/2}(G_n - F_n)$ as $\tilde{B}_n\circ F_n$ and $n^{1/2}(F_n - F)$ as $B_n\circ F$ and call these processes $\tilde{W}_n$ and W_n respectively. Then we can write the right-hand side of (5.3) as

$$-\{(W_n\circ G_n^{-1} - W_n\circ F_n^{-1}) + \tilde{W}_n\circ G_n^{-1}\} - n^{1/2}\{(F_n\circ F_n^{-1} - I) - (G_n\circ G_n^{-1} - I)\}$$

where I is the identify. Therefore, to prove the theorem it is enough to show that the following five assertions, (5.4)–(5.8), hold for almost all $X_1, X_2, \cdots$.

$$\|F_n\circ F_n^{-1} - I\| = o(n^{-1/2}), \tag{5.4}$$

$$n^{1/2}\|G_n\circ G_n^{-1} - I\| \to 0 \tag{5.5}$$

in (conditional) probability,

(5.6) $$\|R_n - f\circ F^{-1}\| \to 0$$

in (conditional) probability,

(5.7) $$-\tilde{W}_n\circ G_n^{-1} \text{ converges weakly to } B, \text{ on } [t_o, t_1]$$

(5.8) $$\| W_n\circ G_n^{-1} - w_n\circ F_n^{-1}\| \to 0$$

in (conditional) probability.

PROOF OF (5.4). F_n has jumps of size $1/n$ only.

PROOF OF (5.5). Bound (5.5) by

$$n^{1/2}\sup_x\{G_n(x+0) - G_n(x)\} \leqq \sup_x|\tilde{W}_n(x+0) - \tilde{W}_n(x)| + n^{-1/2}.$$

Since F is continuous and strictly increasing, so is F^{-1} and

(5.9) $$\sup_x|\tilde{W}_n(x+0) - W_n(x)| = \sup|\tilde{W}_n\circ F^{-1}(x+0) - \tilde{W}_n\circ F^{-1}(x)|.$$

By Theorem 4.1, given $(X_1, \cdots, X_n)$, $\tilde{W}_n\circ F^{-1}$ converge weakly to B which is continuous. Therefore, the expression in (5.9) tends to 0 in conditional probability and (5.5) follows.

PROOF OF (5.6). By Corollary 4.1 since, by hypothesis, F^{-1} is continuous on (0, 1),

(5.10) $$\|G_n^{-1} - F^{-1}\| \to 0$$

in conditional probability, for almost all $X_1, X_2, \cdots$. Similarly, by the Glivenko-Cantelli Theorem, with probability 1,

$$\|F_n^{-1} - F^{-1}\| \to 0.$$

Claim (5.6) follows since the assumed continuity of F on R implies that F is uniformly differentiable on all compact subsets of R.

PROOF OF (5.7). By (5.10) and Theorem 4.1, given $(X_1, \cdots, X_n)$, the processes $(-\tilde{W}_n\circ F^{-1}, F\circ G_n^{-1})$ viewed as probability measures on $D[t_o, t_1] \times D[t_o, t_1]$ converge weakly to (B, I). By the continuity of the composition map $M:(f, g) \to f\circ g$ at all points of $C[0, 1] \times D[t_o, t_1]$, we have $-\tilde{W}_n\circ G_n^{-1}$ converging weakly to B and (5.7) is proven.

PROOF OF (5.8). We have to be careful here to control W_n with probability 1. Since $\|F\circ F_n^{-1} - F\circ G_n^{-1}\| \to 0$ in conditional probability and $W_n = B_n\circ F$, it is enough to check that if $\delta_n \to 0$,

$$\omega(\delta_n, B_n) \to 0 \text{ a.s.}$$

But this follows for instance from Komlos, Major and Tusnady (1975, Theorem 3). The theorem is proved.

REMARKS. (1) If $F^{-1}(0+) > -\infty$ and $F^{-1}(1) < \infty$ and f is continuous on $[F^{-1}(0+), F^{-1}(1)]$, the conclusion of the theorem holds in $D[F^{-1}(0+), F^{-1}(1)]$. For instance, if F is uniform on (0, 1), convergence holds in $D[0, 1]$. More generally, we may have one end of the support finite and the other infinite and have the appropriate theorem hold.

(2) Suppose $\{\tilde{F}_n\}$ is a general sequence of probability measures depending on $X_1, \cdots, X_n$ and G_n is the empirical d.f. of $Y_1, \cdots, Y_n$ which, given $(X_1, \cdots, X_n)$, are i.i.d. with common distribution F_n. We can give simple conditions for $\sqrt{n}\,(G_n^{-1} - \tilde{F}_n^{-1})$ to converge weakly, given $(X_1, \cdots, X_n)$ (as probability measures on $D([t_o, t_1])$ to $B/(f\circ F^{-1})$, provided

that we require the convergence to hold in probability as in Efron. All we need in addition to (5.1) is that (i) $n^{1/2}(\tilde{F}_n - F)$ converge weakly (as probability measures on D) to a limit with continuous sample functions, and (ii) $\sup_x |\tilde{F}_n(x+0) - \tilde{F}_n(x)| = o_p(n^{-1/2})$. Hence the parametric bookstrap works if, for example, $F = F_{\theta_o}$ satisfies (5.1) and $(\partial/\partial\theta) F_\theta|_{\theta_o}$ is continuous in x and $n^{1/2}(\hat{\theta}_n - \theta_o) = O_p(1)$.

Here are some applications which follow fairly easily from the theorem.

The median. Let m^* be the median of the X_ι^* and m the median of the X_ι.

PROPOSITION 5.1. *If F has a unique median μ and f has a positive derivative f continous in a neighborhood of μ, then along almost all sample sequences $X_1, X_2, \cdots$, given $(X_1, \cdots, X_n)$, $n^{1/2}(m^* - m)$ converges weakly to* $N\left(0, \dfrac{1}{4f^2(\mu)}\right)$, *the limit law of* $n^{1/2}(m - \mu)$.

By this result the quantiles of the bootstrap distribution of $n^{1/2}(m^* - m)$ can be used to set an approximate confidence interval for μ. An asymptotic pivot in which we estimate the density f and then scale can also be bootstrapped.

A more careful argument shows that Proposition 5.1 holds under the weakest natural conditions: μ is unique and F has positive derivative f at μ.

Quantile intervals. The usual interval for the population median is $[X_{(k)}, X_{(n-k+1)}]$ where $X_{(1)} < \cdots < X_{(n)}$ are the order statistics of the sample, and k is determined by the desired confidence coefficient through the relation

$$P\{X_{(j)} < \mu \leqq X_{(j+1)}\} = \binom{n}{j} 2^{-n}$$

valid for all continuous F.

Since $X_{(j)} = F_n^{-1}(j/k)$ is the j/k quantile of the law of X_1^*, given $(X_1, \cdots, X_n)$, the bootstrap principle leads us to believe

$$P\{X_{(k)} < M \leqq X_{(\ell)} | F_n\} \approx P\left\{F^{-1}\left(\frac{k}{n}\right) < m \leqq F^{-1}\left(\frac{\ell}{n}\right)\right\} \tag{5.11}$$

where $P(\cdot | F_n)$ is the conditional probability, given $(X_1, \cdots, X_n)$. Efron, by exact calculation, gets the unexpected approximation

$$P\{X_{(k)} < M \leqq X_{(\ell)} | F_n\} \approx P\{X_{(k)} < \mu \leqq X_{(\ell)}\}. \tag{5.12}$$

If we interpret $\approx$ as meaning that the difference of the two sides goes to 0 along almost all sample sequences, then both (5.11) and (5.12) can be established under the assumptions of Theorem 5.1.

Linear combinations of order statistics. Theorem 5.1 establishes the validity of the bootstrap for linear combinations of order statistics with nice weight functions concentrated on $[\alpha, 1-\alpha]$, $0 < \alpha < ½$. That is,

$$n^{1/2}\left\{\int_\alpha^{1-\alpha} F_n^{-1}(t)\, d\Lambda_n(t) - \int_\alpha^{1-\alpha} F^{-1}(t)\, d\Lambda_n(t)\right\}$$

can be bootstrapped under condition (5.1) provided that $\Lambda_n \to \Lambda$ weakly. As a special case, if we take Λ_n to be the uniform distribution on $[\alpha, 1-\alpha]$, we see that the bootstrap provides confidence intervals for the center of symmetry of a symmetric distribution based on the α-trimmed mean. The bootstrap is also valid for estimates of the asymptotic variance of such linear combinations of order statistics and for pivots based on t-like statistics.

6. Counter-examples. In Sections 2 and 3 we checked the validity of the bootstrap for various functionals $R_n\{(X_1, \cdots, X_n); F_n\}$. Roughly, the bootstrap will work provided that

(6.1a) $R_n\{(Y_1, \cdots, Y_n); G\}$ tends weakly to a limit law $\mathscr{L}_G$ whenever $Y_1, \cdots, Y_n$ are i.i.d. with distribution G, for all G in a "neighborhood" of F into which F_n falls eventually with probability 1,

(6.1b) the convergence in (6.1a) is uniform on the neighborhood,

and

(6.1c) the function $G \to \mathscr{L}_G$ is continuous.

In the examples of this section, the bootstrap fails because uniformity does not hold on any usable neighborhoods.

Counter-example 1: a U-statistic. Let

$$R_n(Y_1, \cdots, Y_n; G) = n^{1/2}\left\{\binom{n}{2}^{-1} \sum_{i<j} \left[\omega(Y_i, Y_j) - \int \omega(x, y)\, dG(x)\, dG(y)\right]\right\} \tag{6.2}$$

a normalized centered U-statistic. As we have noted in the previous section, by a theorem of Hoeffding, if

$$\int \omega^2(x, y)\, dF(x)\, dF(y) < \infty, \tag{6.3}$$

then

(6.4) $R_n(X_1, \cdots, X_n; F)$ converges weakly to a $N(0, \sigma^2)$ random variable,

where σ^2 is given by (3.18).

To bootstrap the U-statistic, however, we have to assume not only (6.3) but also the von Mises condition

$$\int \omega(x, x)^2\, dF(x) < \infty \tag{6.5}$$

Absent this condition, the bootstrap can fail: indeed, $|R(X_1^*, \cdots, X_n^*; F_n)|$ can tend to ∞.

Suppose F is the uniform distribution on $(0, 1)$ and write $\omega = \omega_1 + \omega_2$ where $\omega_1(x, y) = \omega(x, y)I(x \neq y)$. Let R_{n1}, R_{n2} be the U-statistics corresponding to ω_1, ω_2 respectively. Then $R_n = R_{n1} + R_{n2}$. If (6.3) holds, by Theorem 3.1, given $(X_1, \cdots, X_n)$, the conditional distribution of $R_{n1}(X_1^*, \cdots, X_n^*; F_n)$ tends weakly to $N(0, \sigma^2)$. An example will be given where $|R_{n2}(X_1^*, \cdots, X_n^*; F_n)|$ tends to ∞ in probability. Of course, $R_{n2}(X_1, \cdots, X_n; F) = 0$.

To develop this example, write

$$R_{n2}(X_1^*, \cdots, X_n^*; F_n) = \{n^{1/2}(n-1)\}^{-1} \sum_{i=1}^n \omega(X_i, X_i)\left\{\nu_{in}(\nu_{in} - 1) - \frac{n-1}{n}\right\}, \tag{6.6}$$

where

(6.7) ν_{in} is the number of j's with $1 \leqq j \leqq n$ and $X_j^* = X_i$.

Let $Z_i = \omega(X_i, X_i)$, $i = 1, \cdots, n$ and $Z_{(1)} \leqq \cdots \leqq Z_{(n)}$ be the corresponding order statistics. Take

$$\omega(x, x) = e^{1/x}.$$

We claim

(6.8) the conditional distribution of $\{n^{1/2}(n - 1/Z_{(n)}\}R_{n2}(X_1^*, \cdots, X_n^*; F_n)$ converges in probability to a limit law, namely the distribution of $\nu(\nu - 1) - 1$ where ν is a Poisson variable with mean 1.

Moreover

(6.9) $n^A/Z_{(n)}$ tends to 0 in probability as $n \to \infty$, for every positive A.
So R_{n2} does indeed dominate R_{n1}.

Our assertions about the behavior of R_n are proved as follows. Let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics of $X_1, \cdots, X_n$. Then the distribution of

$$n^{-1}(\log Z_{(n)} - \log Z_{(n-1)}) = \frac{n(X_{(2)} - X_{(1)})}{(n^2 X_{(1)} X_{(2)})}$$

converges to a limit concentrating on $(0, \infty)$, since $nX_{(1)}$ and $n(X_{(2)} - X_{(1)})$ converge jointly in law to two independent exponentials. Therefore,

(6.10) $n^A Z_{(n-1)}/Z_{(n)}$ tends to 0 in probability, for any positive A.
Let I be the "antirank" of $Z_{(n)}$, defined by $Z_I = Z_{(n)}$. Then,

$$n^{1/2}(n-1)R_{n2}(X_1^*, \cdots, X_n^*; F_n)/Z_{(n)} = \nu_{In}(\nu_{In} - 1) + O_p\{n^2 Z_{(n-1)}/Z_{(n)}\},$$

since $\sum \nu_{in}(\nu_{in} - 1) \leqq n(n-1)$.

Now (6.8) follows: given $X_1, \cdots, X_n$, conditionally ν_{In} has a binomial distribution with n trials and success probability $1/n$, whose limit is Poisson with mean 1. The remainder is negligible, by (6.10).

The claim (6.9) follows by a previous argument, since $n^{-1} \log Z_{(n)} = (nU_{(1)})^{-1}$ converges in law.

Counter-example 2: the maximum and spacings. If F is uniform on $(0, \theta)$, the usual pivot for θ is $n(\theta - X_{(n)})/\theta$ which has a limiting standard exponential distribution. If we think of θ as the upper end point of the support of F then it is natural to bootstrap $(n(\theta - X_{(n)})/\theta$ by $n(X_{(n)} - X_{(n)}^*)$, where $X_{(1)}^* \leqslant \cdots \leqslant X_{(n)}^*$ are the ordered X_i^*. This does not work. In fact,

$$P\{n(X_{(n)} - X_{(n)}^*) = 0 \mid F_n\} \to 1 - e^{-1} \doteq 0.63.$$

More generally, it is easy to see that for almost all $X_1, X_2, \cdots$,

$$P\{X_{(n)}^* < X_{(n-k+1)} \mid F_n\} \to e^{-k}, \qquad k = 1, \cdots.$$

Thus, with probability 1, the conditional distribution of $n(X_{(n)} - X_{(n)}^*)/X_{(n)}$ does not have a weak limit: since $\limsup n(X_{(n)} - X_{(n-k+1)}) = \infty$, and $\liminf n(X_{(n)} - X_{(n-k+1)}) = 0$, a.s. for each k.

This unpleasant behavior cannot be mended by simple smoothing, e.g., replacing F_n by $\tilde{F}_n$ which puts mass $1/(n-1)$ uniformly into each interval $[X_{(n-k+1)}, X_{(n-k)}]$, for $k = 0, \cdots, n-2$. Nor does this behavior have much to do with the maximum. The conditional distributions of the spacings $n(X_{(k)}^* - X_{(k-1)}^*)$ do not have weak limits, even though $n(X_{(k)} - X_{(k-1)})$ has an exponential limit.

The problem is the lack of uniformity in the convergence of F_n to F. Uniformity does hold for the parametric bootstrap, where F is estimated by $\hat{F}_n$, which is uniform on the interval $(0, X_{(n)})$. If $X_1^*, \cdots, X_n^*$ are a sample from $\hat{F}_n$, then

$$\mathscr{L}(X_1^*/X_{(n)}, \cdots, X_n^*/X_{(n)}) = \mathscr{L}(X_1/\theta, \cdots, X_n/\theta)$$

7. Other work. Freedman (1981) has pursued the use of the bootstrap for least squares estimates in regression models when the number of parameters is fixed, and arrived at results very similar to those obtained for means in the one-sample problem. Work is in progress at Berkeley on the behavior of other types of estimates in these models, as well as on the general theory of bootstrapping von Mises functionals in one-sample models.

The authors are studying the behavior of the bootstrap in regression models when the number of parameters is large as well as the sample size; also considered is the sampling of finite populations. An interesting new phenomenon surfaces: the bootstrap can work for linear statistics based on large numbers of summands even though the normal approximation does not hold. On the other hand, the bootstrap fails quite generally when the number of parameters is too large.

8. Mathematical appendix. In Section 2, we used the Mallows metric d_2 and its cousin d_1. It may be helpful to give a fuller account of such metrics here. Let B be a separable Banach space with norm $\|\cdot\|$. The only present case of interest is finite-dimensional Euclidean space, in the Euclidean norm. Let $1 \leqq p < \infty$; only $p = 1$ or 2 are of present interest.[3]

Let $\Gamma_p = \Gamma_p(B)$ be the set of probabilities γ on the Borel σ-field of B, such that $\int \|x\|^p \gamma(dx) < \infty$. For α and β in Γ_p, let $d_p(\alpha, \beta)$ be the infimum of $E\{\|X - Y\|^p\}^{1/p}$ over pairs of B-valued random variables X and Y, where X has law α and Y has law β.

LEMMA 8.1. (a) *The infimum is attained.*
(b) *d_p is a metric on Γ_p.*

PROOF: *Claim* (*a*). Let X and Y be the coordinate functions on $B \times B$. Using weak compactness, it is easy to find a probability π on $B \times B$, such that $\pi X^{-1} = \alpha$, and $\pi Y^{-1} = \beta$, and $\int \|X - Y\|^p d\pi$ is minimal.

Claim (*b*). Only the triangle inequality presents any problem. Fix α, β and γ in Γ_p. Using the first claim, choose π on $B \times B$ so $[\int \|X - Y\|^p d\pi]^{1/p} = d_p(\alpha, \beta)$. Changing notation slightly, let Y and Z be the coordinates on another "plane" $B \times B$; find π' on this $B \times B$ so $[\int \|Y - Z\|^p d\pi']^{1/p} = d_p(\beta, \gamma)$. Now stitch the two planes together along the Y-axis into a 3-space $B \times B \times B$. More formally, let X, Y, Z be the coordinate functions on $B \times B \times B$. Define π^* on $B \times B \times B$ by the requirements:

- the π^*-law of Y is β;
- given Y, the variables X and Z are conditionally π^*-independent;
- the conditional π^*-law of X given $Y = y$ coincides with the conditional π-law of X given $Y = y$;
- the conditional π^*-law of Z given $Y = y$ coincides with the conditional π'-law of Z given $Y = y$.

In particular, the π^*-law of (X, Y) is π; the π^*-law of (Y, Z) *is* π'.

Minkowski's inequality can now be used, as follows:

$$\begin{aligned} d_p(\alpha, \gamma) &\leqq \left\{\int \|X - Z\|^p d\pi^*\right\}^{1/p} \\ &\leqq \left\{\int [\|X - Y\| + \|Y - Z\|]^p d\pi^*\right\}^{1/p} \\ &\leqq \left\{\int \|X - Y\|^p d\pi^*\right\}^{1/p} + \left\{\int \|Y - Z\|^p d\pi^*\right\}^{1/p} \\ &= \left\{\int \|X - Y\|^p d\pi\right\}^{1/p} + \left\{\int \|Y - Z\|^p d\pi'\right\}^{1/p} \\ &= d_p(\alpha, \beta) + d_p(\beta, \gamma) \end{aligned}$$ □

On the real line, Lemma 8.2 below gives a very convenient representation for d_p (see Major, 1978). In this case, the probabilities α and β are defined by their distribution functions F and G.

[3] The essential supremum corresponds to $p = \infty$ and can be handled analogously. The extension to Orlicz spaces might be useful: see Zaanen (1953) or Zygmund (1935).

LEMMA 8.2. *If B is the real line, with $\|x\| = |x|$, then*

$$d_p(F, G) = \left\{\int_0^1 |F^{-1}(t) - G^{-1}(t)|^p \, dt\right\}^{1/p}$$

The case $p = 1$ is especially simple because

$$\int_0^1 |F^{-1}(t) - G^{-1}(t)| \, dt = \int_{-\infty}^{\infty} |F(t) - G(t)| \, dt. \tag{8.1}$$

Indeed, both sides of (8.1) represent the area between the graphs of F and G.

Return now to the general setting.

LEMMA 8.3. *Let $\alpha_n, \alpha \in \Gamma_p$. Then $d_p(\alpha_n, \alpha) \to 0$ as $n \to \infty$ is equivalent to each of the following.*

a) *$\alpha_n \to \alpha$ weakly and $\int \|x\|^p \alpha_n(dx) \to \int \|x\|^p \alpha(dx)$.*
b) *$\alpha_n \to \alpha$ weakly and $\|x\|^p$ is uniformly α_n-integrable.*
c) *$\int \phi \, d\alpha_n \to \int \phi \, d\alpha$ for every continuous ϕ such that $\phi(x) = 0(\|x\|^p)$ at infinity.*

PROOF. a) "Only if". Suppose $d_p(\alpha_n, \alpha) \to 0$. Let ξ_n have law α_n, and ζ have law α, and $E[\|\xi_n - \zeta\|^p]^{1/p} = d_p(\alpha_n, \alpha)$. Then

$$\left[\int \|x\|^p \alpha_n\,(dx)\right]^{1/p} - \left[\int \|x\|^p \alpha\,(dx)\right]^{1/p} = E\{\|\xi_n\|^p\}^{1/p} - E\{\|\xi\|^p\}^{1/p}$$
$$\leqq E\{\|\xi_n - \zeta\|^p\}^{1/p} \to 0$$

Likewise, if f is Lipschitz, that is $\|f(x) - f(y)\| \leqq K\|x - y\|$, then

$$\left|\int f(x)\alpha_n\,(dx) - \int f(x)\alpha\,(dy)\right| = |E\{f(\xi_n) - f(\zeta)\}| \leqq E\{|f(\xi_n) - f(\zeta)|\}$$
$$\leqq KE\{\|\xi_n - \zeta\|\} \leqq KE[\|\xi_n - \zeta\|^p]^{1/p} \to 0.$$

Then $\alpha_n \to \alpha$ weakly by a routine argument.

"If". Suppose $\alpha_n \to \alpha$ weakly and $\int\|x\|^p\alpha_n\,(dx) \to \int\|x\|^p\alpha\,(dx)$. A routine argument reduces the problem to the case where α_n and α concentrate on a fixed bounded set, using the condition on the norms; then the reduction to the case where α_n and α concentrate on a fixed compact set C is easy, using Prokhorov's theorem (Billingsley, 1968, page 37). Cover C by a finite disjoint union of sets C_i of diameter ϵ, with $\alpha(\partial C_i) = 0$, where ∂ represents the boundary. Choose $x_i \in C_i$. Replace α_n by $\tilde{\alpha}_n$, where $\tilde{\alpha}_n\{x_i\} = \alpha_n\{C_i\}$. Likewise for α. Clearly $d_p(\tilde{\alpha}_n, \alpha_n) \leqq \epsilon$ and $d_p(\tilde{\alpha}, \alpha) \leqq \epsilon$. But $d_p(\tilde{\alpha}_n, \tilde{\alpha}) \to 0$ by an easy direct argument. The rest is immediate. □

The argument for the "if" part of (a) is a variation on an argument for Vitali's theorem.

LEMMA 8.4. *Let X_i be independent B-valued random variables, with common distribution $\mu \in \Gamma_p$. Let μ_n be the empirical distribution of $X_1, \cdots, X_n$. Then $d_p(\mu_n, \mu) \to 0$* a.e.

PROOF. Use Lemma 8.3 and the strong law. □

For B-valued random variables U and V, write $d_p(U, V)$ for the d_p-distance between the laws of U and V, assuming the latter are in Γ_p. The scaling properties of d_p are as follows:

$$d_p(aU, aV) = |a| \cdot d_p(U, V) \qquad \text{for any scalar } a \tag{8.2}$$

(8.3) $d_p(LU, LV) \leqq \|L\| \cdot d_p(U, V)$ for any linear operator L on B.

The next lemma involves two separable Banach spaces B and B', e.g., two finite-dimensional Euclidean spaces. Let $1 \leqq p, p' < \infty$.

LEMMA 8.5. *Suppose X_n is a B-valued random variable and $\|X_n\| \in L_p$; likewise for X; and $d_p(X_n, X) \to 0$. Let ϕ be a continuous function from B to B', and $\|\phi(x)\|^{p'} \leqq K\{1 + \|x\|^p\}$, where K is some constant. Then $d_{p'}[\phi(X_n), \phi(X)] \to 0$.*

PROOF. Use Lemma 8.3.

Can $d_{p'}[\phi(X_n), \phi(X)]$ be bounded above by some reasonable function of $d_p(X_n, X)$? Apparently not. Suppose $B = B'$ is the real line, $p = 2$ and $p' = 1$ and $\phi(x) = x^2$. Find real numbers x_n and y_n with $(x_n - y_n)^2 \to 0$ but $|x_n^2 - y_n^2| \to \infty$. Let $X_n = x_n$ and $Y_n = y_n$ a.s. Then $d_2(X_n, Y_n) \to 0$ but $d_1(X_n^2, Y_n^2) \to \infty$.

LEMMA 8.6. *Let U_j be independent; likewise for V_j; assume the laws are in Γ_p. Then*

$$d_p\left(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j\right) \leqq \sum_{j=1}^m d_p(U_j, V_j).$$

PROOF. In view of Lemma 8.1, assume without loss of generality that the pairs (U_j, V_j) are independent and

$$E\{\|U_j - V_j\|^p\}^{1/p} = d_p(U_j, V_j).$$

Now by Minkowski's inequality,

$$\begin{aligned} d_p\left(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j\right) &\leqq E\left\{\left\|\sum_{j=1}^m (U_j - V_j)\right\|^p\right\}^{1/p} \\ &\leqq \sum_{j=1}^m E\{\|U_j - V_j\|^p\}^{1/p} = \sum_{j=1}^m d_p(U_j, V_j). \quad \square \end{aligned}$$

In the presence of orthogonality, this result can be improved.

LEMMA 8.7. *Suppose B is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, and $p = 2$. Suppose the U_j are independent, likewise for V_j; assume the laws are in Γ_2, and $E(U_j) = E(V_j)$. Then*

$$d_2\left(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j\right)^2 \leqq \sum_{j=1}^m d_2(U_j, V_j)^2.$$

PROOF. Make the same construction as in the previous lemma. Now $E\{\langle U_j - V_j, U_k - V_k\rangle\}$ is 0 or $d_2(U_j, V_j)^2$, according as $k \neq j$ or $k = j$. So

$$\begin{aligned} d_2\left(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j\right)^2 &\leqq E\left\{\left\langle \sum_{j=1}^m (U_j - V_j), \sum_{j=1}^m (U_j - V_j)\right\rangle\right\} \\ &= \sum_{j=1}^m d_2(U_j, V_j)^2. \quad \square \end{aligned}$$

LEMMA 8.8. *Suppose B is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, and $p = 2$. Let U and V be B-valued random variables, with $\|U\|$ and $\|V\|$ in L_2. Then*

$$d_2[U, V]^2 = d_2[U - E(U), V - E(V)]^2 + \|E(U) - E(V)\|^2.$$

PROOF. Write $a = E(U)$ and $b = E(V)$. Choose U and V so that $E(\|U - V\|^2) = d_2(U, V)^2$. Now

$$E\{\|(U - a) - (V - b)\|^2\} = E(\|U - V\|^2) - \|a - b\|^2$$

so

$$d_2(U - a, V - b)^2 \leqq d_2(U, V)^2 - \|a - b\|^2.$$

For the other inequality, choose U and V so that

$$E\{\|(U-a)-(V-b)\|^2\}=d_2(U-a,\,V-b)^2. \qquad \square$$

For simplicity, the next result will be given only for the line.

LEMMA 8.9. *Suppose B is the real line, $\|x\| = |x|$, and $p = 2$. Let d_2^1 be the corresponding Mallows metric. Let $U_1, \cdots, U_n$ be independent and identically distributed L_2-variables, and let U be the column vector $(U_1, \cdots, U_n)$. Let $V_1, \cdots, V_n$ and V be likewise. Suppose $E(U_i) = E(V_i)$. Let A be an $m \times n$ matrix of scalars. Now AU, AV are random vectors in R^m, equipped with the m-dimensional Euclidean norm. Write d_2^m for the corresponding d_2-metric. Then*

$$d_2^m(AU,AV)^2 \leqq \operatorname{trace}(AA')\cdot d_2^1(U_i,V_i)^2.$$

PROOF. As usual, suppose (U_i, V_i) are independent and $E\{(U_i - V_i)^2\}^{1/2} = d_2(U_i, V_i)$. Now

$$\begin{aligned} d_2(AU,AV)^2 &\leqq E\{\|AU-AV\|^2\}\\ &= E\{\operatorname{trace}[A(U-V)(U-V)'A']\}\\ &= \operatorname{trace}(AA')\cdot d_2^1(U_i,V_i)^2 \end{aligned}$$

because $E\{(U-V)(U-V)'\} = I_{n\times n}\cdot d_2^1(U_i, V_i)^2$, where $I_{n\times n}$ is the $n \times n$ identity matrix, and trace CD = trace DC, provided both matrix products make sense. $\square$

The next result expresses the idea that the bootstrap operation commutes with smooth functions. Let ϕ be a function from one separable Banach space B to another B'. Let $x_0 \in B$; most of the action will occur near x_0. Suppose that ϕ is continuously differentiable at x_0 in the following sense. For some $\delta_0 > 0$, if $\|x - x_0\| \leqq \delta_0$, then as real $h \to 0$,

$$\frac{\phi(x+hy)-\phi(x)}{h} \to \phi'(x)y \quad \text{weakly}$$

for all $y \in B$, where $\phi'(x)$ is a bounded linear mapping from B to B'. Assume too that if $\|x_n - x_0\| \to 0$ then $\|\phi'(x_n)y - \phi'(x_0)y\| \to 0$, uniformly on strongly compact y-sets. By the uniform boundedness principle, there is a positive $\delta_1 \leqq \delta$ such that $\|x - x_0\| \leqq \delta_1$ entails $\|\phi'(x)\| \leqq K$.

LEMMA 8.10. *Let X_n be a B-valued random variable and a_n a scalar tending to infinity, and $x_n \in B$ with $x_n \to x_0$. Suppose the law of $a_n(X_n - x_n)$ converges weakly to the law of W. Let ϕ be a smooth function from B to B', as above. Then the law of $a_n[\phi(X_n) - \phi(x_n)]$ converges weakly to the law of $\phi'(x_0)W$.*

PROOF. The argument is only sketched. Fix a bounded linear functional λ on B, an $x \in B$ with $\|x - x_0\| < \frac{1}{2}\delta_1$, a $y \in B$ with $\|y\| < \frac{1}{2}\delta_1$, and let t be real with $|t| \leqq 1$. Then

$$\frac{\partial}{\partial t}\lambda[\phi(x+ty)] = \lambda[\phi'(x+ty)y]. \tag{8.4}$$

The right hand side of (8.4) is a bounded function of t, so $t \to \lambda[\phi(x + ty)]$ is absolutely continuous, and

$$\lambda[\phi(x+ty)] = \lambda[\phi(x)] + \int_0^t \lambda[\phi'(x+uy)y]\,du. \tag{8.5}$$

Since (8.5) holds for all λ,

$$\phi(x+ty) = \phi(x) + \int_0^t \phi'(x+uy)y\,du \tag{8.6}$$

where $u \to \phi'(x + uy)y$ is strongly integrable by a direct argument. If n is large, $\|x_n - x_0\| < \frac{1}{2}\delta_1$; and $\|X_n - x_n\| < \frac{1}{2}\delta_1$ with overwhelming probability. Then, except for a set of uniformly small probability, by substitution into (8.6),

$$a_n[\phi(X_n) - \phi(x_n)] = \int_0^1 \phi'[x_n + u(X_n - x_n)]\, a_n(X_n - x_n)\, du. \tag{8.7}$$

By Prokhorov's theorem, except on a set of uniformly small probability, $a_n(X_n - x_n) \in C$, a fixed large compact set. So, except for a set of uniformly small probability, the integrand on the right is uniformly close to $\phi'(x_0)a_n(X_n - x_n)$; this final approximation is even uniform in u. □

REMARK. The interaction of two standard terminologies is perhaps unfortunate: if b_n and $b \in B$, then $b_n \to b$ weakly means $\lambda(b_n) \to \lambda(b)$ for all bounded linear functionals λ on B. On the other hand, if W_n and W are B-valued random variables, the law of W_n converges weakly to the law of W iff $E\{\theta(W_n)\} \to E\{\theta(W)\}$ for all bounded functions θ on B which are continuous in the strong topology.

LEMMA 8.11. *If B is the Brownian bridge and T is a closed subset of* $[0, 1]$ *which contains points other than* 0 *and* 1, *then* $\sup_T|B(t)|$ *has a continuous distribution.*

Much more is probably true. The distribution of $\sup_T|B(t)|$ may well have a $\mathbb{C}^\infty$ density, and likewise for other diffusions. However, Lemma 8.11 is all we need for Corollary 4.2. To prove the lemma we need a couple of sub-lemmas. Recall that $B(\cdot)$ is a continuous Markov process.

LEMMA 8.11.1. *Let $\mathfrak{B}(t+)$ be the σ field in $\mathbb{C}[0, 1]$ of events which depend only on path behavior right after t* (Freedman, 1971, page 102). *Let P be the probability measure on $\mathbb{C}[0, 1]$ which makes the coordinate process a Brownian bridge. $\mathfrak{B}(t+)$ is trivial,* i.e., *if $A \in \mathfrak{B}(t+)$, then the conditional probability*

$$P(B \in A \mid B(t)) = 0 \quad \text{or} \quad 1$$

with probability 1.

PROOF. Given $B(t) = c$, the process $B(t + u)$ *for* $0 \leqq u \leqq 1 - t$ is Gaussian with the same joint distribution as

$$\sqrt{1-t}\, B\left(\frac{\tau}{1-t}\right) + c\frac{(1-t-\tau)}{1-t}.$$

By a remark of Doob (1949) this in turn has the same joint distributions as

$$\sqrt{1-t}\left(1 - \frac{u}{1-t}\right) W\left(\frac{u}{1-t-u}\right) + c\frac{(1-t-u)}{1-t}$$

where W is a Wiener process on $(0, \infty)$ and $W(0) = 0$. Lemma 8.11.1 follows from the Blumenthal $0 - 1$ law (see Freedman, 1971, page 106, for example).

LEMMA 8.11.2. *We can represent T as the union of two sets, T_{12} and $T - T_{12}$, such that every point in T_{12} may be approached by other points in T from both sides and $T - T_{12}$ is countable.*

PROOF. We can write $T = T_1 \cup T_2$ where T_1 is a closed perfect set and T_2 is countable (Hausdorff, 1957, page 159). Call a point of T_1 an endpoint if it can only be approached on one side by points in T_1. The set of endpoints, call it T_{11}, is clearly countable. Write $T_{12} = T_1 - T_{11}$.

PROOF OF LEMMA 8.11. Note that $\sup_T|B(t)|$ is actually a maximum since B is continuous and, moreover, that $\max_T|B(t)| > 0$ with probability 1 since T includes points other than $\{0, 1\}$. So what we need to prove is, for each $c > 0$,

$$P[\max_T|B(t)| = c] = 0.$$

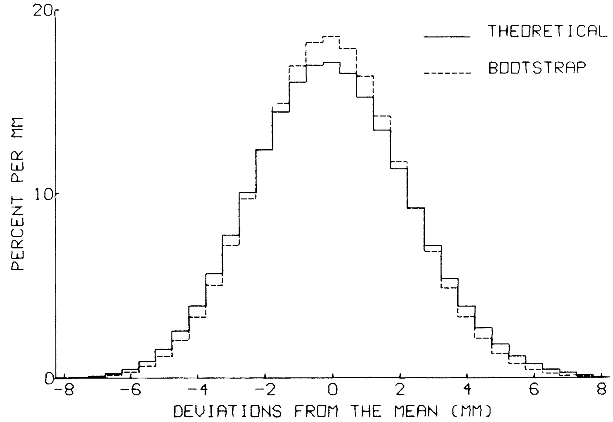


FIG. 1

A simulation, in which the bootstrap distribution is compared to the theoretical distribution.

We claim it is enough to show

(8.8) $$P[\max_{T_{12}} |B(t)| = c, |B(t)| < c : t \in T - T_{12}] = 0$$

since for $c > 0$,

(8.9) $$\sum\{P[|B(t)| = c] : t \in T - T_{12}\} = 0.$$

Associate with each $t \in T_{12}$ in a measurable way a decreasing sequence $s_n(t) \downarrow t$, $s_n(t) \in T \;\forall\; n, t$. For example, take $s_n(t)$ to be the largest point in T which lies between t and $t + 1/n$. Now let σ be the first $t \in T$ such that $|B(t)| = c$ and $\sigma = 1$ otherwise. Then,

(8.10) $$P[\max_{T_{12}} |B(t)| = c, |B(t)| < c, t \in T - T_{12}] \leqq P[\sigma \in T_{12}, |B(s_n(\sigma))| < |B(\sigma)| \text{ for large } n].$$

But by Lemma 8.11.1, for any $t \in T_{12}$

(8.11) $$P[|B(s_n(t))| < |B(t)| \text{ for large } n \,|\, B(t)] = 0 \text{ or } 1.$$

Since $t \in T_{12}$, $\liminf_n P[|B(s_n(t))| \geqq |c| \; |B(t) = c] > 0$ for any finite c and hence the probability in (8.11) is 0. By the strong Markov property the right-hand side of (8.10) is 0. Then (8.8) and the lemma follow. □

9. A simulation. To illustrate Theorem 1.1, a simulation was performed. The population consisted of the 6,672 Americans aged 18–79 in Cycle I of the Health Examination Survey.[4] The variable of interest was systolic blood pressure, with an average of 130.3 and a SD of 23.2 millimeters of mercury. The distribution had a longish right tail: the minimum was 73, the maximum 260, with skewness of 1.3 and kurtosis of 2.4.

A sample of 100 was drawn at random, with replacement. The sample average systolic blood pressure was 129.6 with a SD of 21.4. Consider these sample results from the point of view of a statistician who does not know the population figures, and has forgotten the "SD/$\sqrt{n}$" formula. Such a statistician could estimate the sampling error in the sample

[4] These 6,672 subjects were themselves a probability sample drawn from the American population. The data were provided by the National Center for Health Statistics.

average by the bootstrap principle (Theorem 1.1). The sampling error follows the theoretical sampling distribution of

$$\frac{X_1 + \cdots + X_{100}}{100} - \mu$$

where X_i is the blood pressure of the ith sample subject, and μ is the population average. This is approximated by the bootstrap distribution of

$$\frac{X_1^* + \cdots + X_{100}^*}{100} - \frac{X_1 + \cdots + X_{100}}{100},$$

where the X_i^* are drawn at random with replacement from $\{X_1, \cdots, X_{100}\}$, conditioning on these original X's.

Figure 1 compares the bootstrap distribution (dashed) with the theoretical distribution (solid). Both are rescaled convolutions, one of the population distribution, the other of the sample empirical distribution. These convolutions were computed exactly, using an algorithm based on the Fast Fourier Transform. As the figure shows, the bootstrap distribution follows the theoretical distribution rather closely.

Acknowledgment. We thank Persi Diaconis and Brad Efron for a number of helpful conversations.

REFERENCES

BERK, R. H. (1966). Limiting behaviour of posterior distributions when the model is incorrect. *Ann. Math. Statist.* **37** 51–58.

BICKEL, P. J. (1966). Some contributions to the theory of order statistics. *Proceedings Fifth Berkeley Symp. Math. Statist. and Prob.* **1** 575–592.

BICKEL, P. J. and FREEDMAN, D. A. (1981). More on bootstrapping regression models. Technical report, Statistics Department, University of California, Berkeley.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

DOBRUSHIN, R. L. (1970). Describing a system of random variables by conditional distributions. *Theory Probab. Appl.* **15** 458–486 [especially Section 3].

DOOB, J. L. (1949). Heuristic approach to the Kolmogorov Smirnov theorems. *Ann. Math. Statist.* **20** 393–403.

EFRON, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7** 1–26.

FILLIPOVA, A. R. (1962). Mises' theorem on the asymptotic behavior of functionals of empirical distribution functions and its statistical applications. *Theory Probab. Appl.* **7** 24–57.

FREEDMAN, D. A. (1971). *Brownian Motion and Diffusion.* Holden-Day, San Francisco.

FREEDMAN, D. A. (1981). Bootstrapping regression models. *Ann. Statist.*, **9** 1218–1228.

HAUSDORFF, F. (1957). *Set Theory.* Chelsea, New York.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.* **19** 293–325.

KOMLOS, J., MAJOR, P. and TUSNADY, G. (1975). An approximation of partial sums of independent r.v.'s and the sample d.f. I. *Z. Warsch. verw. Gebiete* **32** 111–131.

MAJOR, P. (1978). On the invariance principle for sums of independent, identically distributed random variables. *Jour. of Multivariate Anal.* **8** 487–501.

MALLOWS, C. L. (1972). A note on asymptotic joint normality. *Ann. Math. Statist.* **43** 508–515.

MISES, R. VON (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* **18** 309–348.

PYKE, R. and SHORACK, G. (1968). Weak convergence of a two-sample empirical process and a new approach to the Chernoff-Savage theorems. *Ann. Math. Statist.* **39** 755–771.

REEDS, J. (1976). On the definition of von Mises functionals. Thesis, Harvard University.

SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187–1195.

STIGLER, S. (1973). The asymptotic distribution of the trimmed mean. *Ann. Statist.* **6** 472–477.

TANAKA H. (1973). An inequality for a functional of probability distribution and its application to Kac's one-dimensional model of a Maxwellian gas. *Z. Warsch. verw. Gebiete* **27** 47–52.

VALLENDER, S. S. (1973). Calculation of the Vasershtein distance between probability distributions on the line. *Theory Probab. Appl.* **18** 784–786.

ZAANEN, A. C. (1953). *Linear Analysis.* Wiley, New York.

ZYGMUND, A. (1935). *Trigonometric Series*, (reprinted by Dover and by Chelsea, New York).

Statistics Department
University of California, Berkeley
BERKELEY, CALIFORNIA 94720

The Annals of Statistics
1984, Vol. 12, No. 2, 470–482

ASYMPTOTIC NORMALITY AND THE BOOTSTRAP IN STRATIFIED SAMPLING

BY P. J. BICKEL[1] AND D. A. FREEDMAN[2]

University of California, Berkeley

This paper is about the asymptotic distribution of linear combinations of stratum means in stratified sampling, with and without replacement. Both the number of strata and their size is arbitrary. Lindeberg conditions are shown to guarantee asymptotic normality and consistency of variance estimators. The same conditions also guarantee the validity of the bootstrap approximation for the distribution of the t-statistic. Via a bound on the Mallows distance, situations will be identified in which the bootstrap approximation works even though the normal approximation fails. Without proper scaling, the naive bootstrap fails.

1. Introduction. Consider the problem of estimating a linear combination $\gamma = \sum_{i=1}^{p} c_i\mu_i$ of the means $\mu_1, \cdots, \mu_p$ of p numerical populations $X_1, \cdots, X_p$ with corresponding distributions $F_1, \cdots, F_p$. For each $i = 1, \cdots, p$ there is a sample X_{ij} from population $\mathscr{X}_i$; the sample elements are indexed by $j = 1, \cdots, n_i$. Thus, n_i is the size of the sample from the ith population. Two situations will be discussed:

(a) The populations $\mathscr{X}_i$ are assumed arbitrary and the sampling is with replacement: X_{ij} for $j = 1, \cdots, n_i$ are identically distributed with common distribution F_i; all the X_{ij} are independent.

(b) The populations are assumed finite; $\mathscr{X}_i$ has known size N_i; sampling is without replacement and independent in i; in this case, F_i is uniform. Enumerate X_i as $\{x_{i1}, \cdots, x_{iN_i}\}$.

For simplicity, the populations are supposed univariate.

The natural unbiased estimate of γ is

$$\hat{\gamma} = \textstyle\sum_{i=1}^{p} c_i X_{i\cdot}\,. \tag{1}$$

Here, the dot is the averaging operator.

Let τ_a^2 or τ_b^2 denote the variance of $\hat{\gamma}$ under sampling schemes (a) and (b) respectively. Let $\hat{\tau}_a^2$ or $\hat{\tau}_b^2$ be the customary unbiased variance estimates. Inference about γ can be based either on the normal approximation to the distribution of $(\hat{\gamma} - \gamma)/\hat{\tau}$ or on bootstrap approximations. This paper will discuss the validity of these approximations when the total sample size tends to ∞ in any way

Received February 1983; revised December 1983.

[1] This work was performed with the partial support of Office of Naval Research Contract N00014-80-C-0163. The hospitality of the Hebrew University, Jerusalem is also gratefully acknowledged.

[2] Research partially supported by National Science Foundation Grant MCS80-02535.

AMS 1980 *subject classification.* Primary 60F05; secondary 62E20.

Key words and phrases. Bootstrap, asymptotic normality, stratified sampling, standard errors.

whatsoever, e.g., many small samples or a few large samples or some combination thereof. More precisely: suppose p, the c_i, the populations, the N_i, and n_i all depend on an index ν such that $n(\nu) = n_1(\nu) + \cdots + n_p(\nu) \to \infty$ as $\nu \to \infty$. This index will be suppressed in the sequel.

Here are two examples.

(a) The X_{ij} are unbiased measurements of the same quantity μ, taken with p different instruments. So the precision of X_{ij}, viz.,

$$\sigma_i^2 = \int (x - \mu)^2 \, dF_i(x)$$

depends on i. If σ_i^2 is known to be proportional to r_i, then

$$\hat{\gamma} = \sum \frac{n_i}{r_i} X_{i\cdot} / \sum \frac{n_i}{r_i}$$

is the natural estimate of μ.

(b) In the classical stratified sampling model a population $\mathscr{X}$ of size N is broken up into disjoint strata $\mathscr{X}_1, \cdots, \mathscr{X}_p$ of sizes $N_1, \cdots, N_p$ respectively; $\sum_{i=1}^p N_i = N$. From stratum i the sample X_{ij} for $j = 1, \cdots, n_i$ is taken without replacement. Enumerate the ith stratum as $\{x_{i1}, \cdots, x_{iN_i}\}$. The population mean is

$$\gamma = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^{N_i} x_{ij} = \sum_{i=1}^p N_i x_{i\cdot}/N$$

and $\hat{\gamma} = \sum_{i=1}^p N_i X_{i\cdot}/N$ is the usual estimate of γ.

We first take up the normal approximation in case (a). Suppose

$$\int x^2 \, dF_i < \infty \quad \text{and} \quad n_i \geq 2 \quad \text{for} \quad i = 1, \cdots, p. \tag{2}$$

Then

$$\tau_a^2 = \sum_{i=1}^p c_i^2 \sigma_i^2 / n_i \quad \text{where} \quad \sigma_i^2 = \operatorname{var} X_{ij}$$

and

$$\hat{\tau}_a^2 = \sum_{i=1}^p c_i^2 s_i^2 / n_i$$

where

$$s_i^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (X_{ij} - X_{i\cdot})^2.$$

Let

$$\begin{aligned} \phi(x, \varepsilon) &= x \quad \text{for} \quad |x| \geq \varepsilon \\ &= 0 \quad \text{otherwise} \\ \bar{\phi}(x, \varepsilon) &= x - \phi(x, \varepsilon). \end{aligned}$$

Suppose that for all $\varepsilon > 0$,

$$\tau_i^{-2} \sum_{i=1}^{p} n_i^{-1} c_i^2 E\{\phi^2(X_{ij} - \mu_i, \varepsilon n_i \tau_a \,|\, c_i\,|^{-1})\} \to 0. \tag{3}$$

By the Lindeberg-Feller theorem, $(\hat{\gamma} - \gamma)/\tau_a$ converges in law to $\mathscr{N}(0, 1)$, the standard normal distribution.

According to the first main theorem of this paper, conditions (2) and (3) are also sufficient to guarantee that $\hat{\tau}_a^2$ has the right limiting behavior. However, before giving a precise statement, it may be helpful to reformulate condition (3). Let $Y_{ij} = (X_{ij} - \mu_i)/\sigma_i$. Define the "variance weight" of the ith stratum by

$$w_i^2 = c_i^2 \sigma_i^2 / n_i \tau_a^2 = \text{var}\,\{c_i X_{i\cdot}/\tau_a\}.$$

Clearly,

$$\sum_{i=1}^{p} w_i^2 = 1.$$

Condition (3) can then be written

$$\sum_{i=1}^{p} E\{\phi^2(w_i Y_{ij}, \varepsilon\sqrt{n_i})\} \to 0 \quad \text{for all} \quad \varepsilon > 0. \tag{4}$$

THEOREM 1. *If* (2) *and* (4) *hold in case* (a), *then* $\hat{\tau}_a^2/\tau_a^2 \to 1$ *in probability.*

The proof is deferred.

COROLLARY. $(\hat{\gamma} - \gamma)/\hat{\tau}_a$ *tends to* $\mathscr{N}(0, 1)$ *in law.*

We consider next the bootstrap approximation in case (a); also see Babu and Singh (1983). For $i = 1, \cdots, p$, let $\hat{F}_i$ be the empirical distribution of X_{ij} for $j = 1, \cdots, n_i$. Take samples of size n_i with replacement from $\hat{F}_i$. That is, let $\{X_{ij}^*\}$ be conditionally independent given $\mathscr{F}$, the σ-field spanned by $\{X_{ij}\}$; let X_{ij}^* have common distribution $\hat{F}_i$ for $j = 1, \cdots, n_i$. Let

$$\hat{\gamma}^* = \sum_{i=1}^{p} c_i X_{i\cdot}^*, \quad s_i^{*2} = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (X_{ij}^* - X_{i\cdot}^*)^2$$

$$\hat{\tau}_a^{*2} = \sum_{i=1}^{p} c_i^2 s_i^{*2}/n_i, \quad \tilde{\tau}_a^2 = \sum_{i=1}^{p} c_i^2 (n_i - 1) s_i^2/n_i^2.$$

THEOREM 2. *If* (2) *and* (4) *hold in case* (a), *then the conditional distribution of* $(\hat{\gamma}^* - \hat{\gamma})/\tilde{\tau}_a$ *converges weakly to* $\mathscr{N}(0, 1)$ *in probability, and* $\hat{\tau}_a^*/\tilde{\tau}_a$ *converges to* 1 *in probability.*

The proof is deferred. The theorem points to a problem in using the bootstrap to make inferences: the scaling may go wrong. This is because $X_{i\cdot}^*$ has variance $(n_i - 1)s_i^2/n_i^2$, not s_i^2/n_i. To fix ideas, suppose there are many small strata: more particularly, that $n_i \leq k$ for all i. Now

$$\tilde{\tau}_a^2 \leq (k - 1)/k \cdot \hat{\tau}_a^2 \approx (k - 1)/k \cdot \tau_a^2.$$

The bootstrap distribution of $\hat{\gamma}^* - \hat{\gamma}$ has asymptotic scale $\tilde{\tau}_a$, while $\hat{\gamma} - \gamma$ has the scale τ_a.

We take up next the normal approximation in case (b). Suppose

$$2 \le n_i \le N_i - 1. \tag{5}$$

Then

$$\tau_b^2 = \sum_{i=1}^p c_i^2 \frac{\sigma_i^2}{n_i} \frac{(N_i - n_i)}{N_i - 1}$$

and

$$\hat{\tau}_b^2 = \sum_{i=1}^p c_i^2 \frac{s_i^2}{n_i} \frac{(N_i - n_i)}{N_i}.$$

To state the regularity condition, let v_i^2 be the "variance weight" in case (b): $v_i^2 = c_i^2\sigma_i^2(N_i - n_i)/n_i\tau_b^2(N_i - 1) = \operatorname{var}\{c_i X_{i\cdot}/\tau_b\}$. Let ρ_i be "the effective sample size:" $\rho_i = n_i(N_i - 1)/(N_i - n_i)$. Let $\mathscr{Y}_i = \{y_{i1}, \cdots, y_{iN_i}\}$ where $y_{ij} = (x_{ij} - \mu_{i\cdot})/\sigma_i$ and $\sigma_i^2 = N_i^{-1} \sum_{j=1}^{N_i} (x_{ij} - \mu_{i\cdot})^2$. So $Y_{ij} = (X_{ij} - \mu_i)/\sigma_i$ are sampled from $\mathscr{Y}_i$.

The condition is

$$\sum_{i=1}^p N_i^{-1} \sum_{j=1}^{N_i} \phi^2(v_i y_{ij}, \varepsilon\sqrt{\rho_i}) \to 0. \tag{6}$$

This may be compared with condition (4).

If $\sup_{1\le i\le p} E\,|\,Y_i\,|^3$ is uniformly bounded independent of the hidden index ν, the Lindeberg conditions (4) and (6) are implied respectively by the natural conditions $\max_i w_i/\sqrt{n_i} \to 0$ or $\max_i v_i/\sqrt{\rho_i} \to 0$. Thus if the standardized populations have reasonably light tails, asymptotic normality holds if for each stratum the variance weight contribution is small or the stratum is heavily sampled.

THEOREM 3. *If* (5) *and* (6) *hold in case* (b), *then*

i) $$(\hat{\gamma} - \gamma)/\tau_b \to \mathscr{N}(0, 1) \quad \textit{in law}$$

and

ii) $$\hat{\tau}_b/\tau_b \to 1 \quad \textit{in probability.}$$

The proof is deferred.

COROLLARY. $(\hat{\gamma} - \gamma)/\hat{\tau}_b \to \mathscr{N}(0, 1)$ *in law.*

Finally, we consider the bootstrap in case (b). If $N_i/n_i = k_i$ an integer for each i, the natural bootstrap procedure was suggested by Gross (1980): given $\{X_{ij}\}$, to create populations $\hat{\mathscr{X}}_i$ consisting of k_i copies of each X_{ij} for $j = 1, \cdots, n_i$, then X_{ij}^* for $j = 1, \cdots, n_i$ are generated as a sample without replacement from $\hat{\mathscr{X}}_i$, the samples being independent for different $i = 1, \cdots, p$. In general, if $N_i = k_i n_i + r_i$ with $0 \le r_i < n_i$, form populations $\hat{\mathscr{X}}_{i0}$ and $\hat{\mathscr{X}}_{i1}$, where $\hat{\mathscr{X}}_{i0}$ consists of k_i

copies of each X_{ij}, for $j = 1, \cdots, n_i$; while $\mathscr{X}_{i1}$ consists of $k_i + 1$ copies. Let

$$\alpha_i = \left(1 - \frac{r_i}{n_i}\right)\left(1 - \frac{r_i}{N_i - 1}\right).$$

With probability α_i, let $(X_{i1}^*, \cdots, X_{in_i}^*)$ be a sample without replacement of size n_i from $\hat{\mathscr{X}}_{i0}$; with probability $1 - \alpha_i$, let $(X_{i1}^*, \cdots, X_{in_i}^*)$ be a sample without replacement of size n_i from $\hat{\mathscr{X}}_{i1}$. The virtue of this scheme is that both $\hat{\mathscr{X}}_{i0}$ and $\hat{\mathscr{X}}_{i1}$ have the same distribution $\hat{F}_i$ and

$$\operatorname{Var}(X_{i.}^* \mid \{X_{ij}\}) = \frac{n_i - 1}{n_i^2} s_i^2 \left(\frac{N_i - n_i}{N_i - 1}\right).$$

The proof of the following theorem is similar to that of Theorem 2 and is omitted. Define $\hat{\gamma}^*$ as before, and $\hat{\tau}_b^{*2}$ by substituting X_{ij}^* for X_{ij} in $\hat{\tau}_b^2$.

THEOREM 4. *Let $\tilde{\tau}_b^2$ be the variance of $\hat{\gamma}^*$ given the data. Then, if* (5) *and* (6) *hold in case* (b), *the conditional distribution of $(\hat{\gamma}^* - \hat{\gamma})/\tilde{\tau}_b$ converges weakly to $\mathscr{N}(0, 1)$ and $\hat{\tau}_b^*/\tilde{\tau}_b \to 1$ in probability.*

The same inference problem arises as in the case of Theorem 2. The variance of $\hat{\gamma}^*$ given the data is an inconsistent estimate of the variance of $\hat{\gamma}$. We have side-stepped the issue by computing the scale externally to the bootstrap process. Other patches could be made: one is to rescale the elements of $\mathscr{X}_i$; another is to adjust the constants c_i. These fixes are all a bit *ad hoc*.

If γ stays bounded as $\nu \to \infty$, our results extend easily to pivots

$$\frac{g(\hat{\gamma}) - g(\gamma)}{g'(\gamma)\hat{\tau}_b}$$

where g is nonlinear continuously differentiable. The same issue as before arises a fortiori for nonlinear functions. Neither the variance of $g(\hat{\gamma}^*)$ given the data nor its natural approximation $[g'(\hat{\gamma})]^2\tilde{\tau}_b^2$ are consistent estimates of the asymptotic variance of $g(\hat{\gamma})$. A fix which works if $\sum_{i=1}^p |c_i\mu_i|$ stays bounded is as before to rescale the elements of $\mathscr{X}_i$ or the c_i before applying the bootstrap. Alternatives (the jackknife, linearization, BRR) are discussed in Krewski and Rao (1981). For the case of one stratum, Theorem 4 was derived independently by Chao and Lo (1983).

The bootstrap can work even when Theorem 4 fails but the circumstances are artificial. Suppose we have only one stratum and $N_1 - n_1 = k$ for all ν i.e., all but k members are sampled. Since $\sum_{j=1}^{N_1} (x_{1j} - \mu_1) = 0$, the pivot $(\hat{\gamma} - \gamma)/\tau_b$ is distributed as the standardized mean of a sample of size k taken without replacement from the population $\mathscr{Y}_1$. No matter how large N_1 is, if k is small and $\mathscr{Y}_1$ nonnormal, we would not expect the normal approximation to apply to $\hat{\gamma}$. To be specific let F_ν be the uniform distribution on $\mathscr{Y}_1$ and suppose

(7) $\qquad F_\nu$ converges to F in the Mallows d_2-metric,

i.e., $F_\nu \to F$ weakly and $\int x^2\, dF_\nu \to \int x^2\, dF$. Then $(\hat{\gamma} - \gamma)/\tau_b$ is distributed in the limit as the standardized mean of k independent variables identically distributed according to F. On the other hand, since we have sampled nearly the whole population we expect the bootstrap to work.

THEOREM 5. *If* (7) *holds, the conditional distribution of* $(\hat{\gamma}^* - \hat{\gamma})/\tilde{\tau}_b$ *converges weakly in probability to the same limit as that of the unconditional distribution of* $(\hat{\gamma} - \gamma)/\tau_b$. *Moreover,* $\hat{\tau}_b/\tau_b$ *and* $\hat{\tau}_b^*/\tilde{\tau}_b$ *both tend to* 1 *in probability.*

We can extend this result somewhat by replacing (7) with a compactness-in-d_2 condition on $\{F_\nu\}$

$$\limsup_{m\to\infty} \limsup_\nu N_1^{-1} \sum_{j=1}^{N_1} \phi^2(v_1 y_{1j}, m) = 0.$$

This condition is evidently weaker than (6) for $p = 1$. The conclusion now is that the d_2-distance between the conditional distribution of $(\hat{\gamma}^* - \hat{\gamma})/\hat{\tau}_b^*$ and the unconditional distribution of $(\hat{\gamma} - \gamma)/\hat{\tau}_b$ tends in probability to 0. A further extension to an arbitrary number of strata which includes both Theorems 4 and 5 is also possible but not worthwhile.

2. Some lemmas. Recall the truncation operator ϕ from Section 1.

LEMMA 1. a) $$\left| \phi\left(\frac{1}{k} \sum_{i=1}^k y_i, \varepsilon\right) \right| \le \sum_{i=1}^k |\phi(y_i, \varepsilon/k)|; \quad \textit{equivalently,}$$

$$|\phi(\textstyle\sum_1^k y_i, \varepsilon)| \le k \sum_{i=1}^k |\phi(y_i, \varepsilon/k^2)|$$

b) *Let* $Y_1, Y_2, \cdots$ *be independent and identically distributed. Then*

$$E\left\{\phi^2\left(\frac{1}{k} \sum_{i=1}^k Y_i, \varepsilon\right)\right\} \le k^2 E\{\phi^2(Y_i, \varepsilon/k)\}.$$

PROOF. Claim a). As is easily verified,

$$\left| \phi\left(\frac{1}{k} \sum_{i=1}^k y_i, \varepsilon\right) \right| \le \phi\left(\frac{1}{k} \sum_{i=1}^k |y_i|, \varepsilon\right).$$

Without loss of generality, suppose all $y_i \ge 0$. Let $y_{(k)}$ be the largest y_i. If $y_{(k)} < \varepsilon/k$, both sides of the inequality vanish. If $y_{(k)} \ge \varepsilon/k$, the left side is the average of the y_i, or zero; the right side is at least the maximum $y_{(k)}$.

Claim b) follows by the Cauchy-Schwarz inequality. □

LEMMA 2. *Let* $(X_1', \cdots, X_n')$ *and* $(X_1, \cdots, X_n)$ *be distributed respectively as samples with and without replacement from a finite population. Then*

$$E\{\phi^2(\textstyle\sum_{i=1}^n X_i, \varepsilon)\} \le E\{\phi^2(\sum_{i=1}^n X_i', \frac{1}{2}\varepsilon)\}.$$

PROOF. By a theorem of Hoeffding (1963), if g is convex, then

$$E\{g(\textstyle\sum X_i)\} \le E\{g(\sum X_i')\}.$$

Let

$$\begin{aligned} g(x, \varepsilon) &= x^2 && \text{for} \quad |x| \geq \varepsilon \\ &= 2\varepsilon|x| - \varepsilon^2 && \text{for} \quad \tfrac{1}{2}\varepsilon \leq |x| \leq \varepsilon \\ &= 0 && \text{otherwise.} \end{aligned}$$

Then g is convex and

$$\phi^2(x, \varepsilon) \leq g(x, \varepsilon) \leq \phi^2(x, \tfrac{1}{2}\varepsilon).$$

So

$$E\{\phi^2(\sum X_i, \varepsilon)\} \leq E\{g(\sum X_i, \varepsilon)\} \leq E\{g(\sum X_i', \varepsilon)\} \leq E\{\phi^2(\sum X_i', \tfrac{1}{2}\varepsilon)\}. \quad \Box$$

The next result involves the Mallows metric d_2; see Mallows (1972) or Bickel and Freedman (1981).

LEMMA 3. *Let $\mathscr{X}$ and $\mathscr{Y}$ be two finite populations of real numbers, of the same size N. Let F and G be the uniform distributions on $\mathscr{X}$ and $\mathscr{Y}$. Suppose F and G have the same means. Let $X_1, \cdots, X_n$ be a sample of size n, drawn at random without replacement from $\mathscr{X}$; let $F_{(n)}$ be the law of $X_1 + \cdots + X_n$. Likewise for $Y_1, \cdots, Y_n$ and $G_{(n)}$. Then*

$$d_2[F_{(n)}, G_{(n)}]^2 \leq \frac{n(N-n)}{N-1} d_2(F, G)^2.$$

PROOF. Enumerate $\mathscr{X}$ as $x_1 \leq x_2 \leq \cdots \leq x_N$ and $\mathscr{Y}$ as $y_1 \leq y_2 \leq \cdots \leq y_N$. Then

$$(1/N) \sum_{i=1}^N (x_i - y_i)^2 = d_2(F, G)^2.$$

This follows from Bickel and Freedman (1981, Lemmas 8.2 and 8.3). Let $Z = \{1, \cdots, N\}$. Let $Z_1, \cdots, Z_n$ be a sample of size n, drawn at random without replacement from Z. Set $X_i = x_{Z_i}$ and $Y_i = y_{Z_i}$. Now

$$\begin{aligned} d_2[F_{(n)}, G_{(n)}]^2 \leq E\{[\textstyle\sum_{i=1}^n (X_i - Y_i)]^2\} &= \frac{n(N-n)}{N-1} E[(X_i - Y_i)^2] \\ &= \frac{n(N-n)}{N-1} d_2(F, G)^2. \quad \Box \end{aligned}$$

Here is an easy generalization of Lemma 3.

LEMMA 4. *For $i = 0, 1$ let $\mathscr{X}_i = \{x_{i1}, \cdots, \mathscr{X}_{iN_i}\}$ be finite populations and F_i the associated uniform distributions on $\mathscr{X}_i$. Let F_{ni} be the distribution of $\sum_{j=1}^n X_j$ when $X_1, \cdots, X_n$ is a sample without replacement from $\mathscr{X}_i$. Let $n \leq N_0 \leq N_1$. If J is a subset of $\{1, \cdots, N_1\}$, let F_{1J} be the uniform distribution on $\{x_{1j}: j \in J\}$.*

Then,

$$d_2(F_{n0}, F_{n1})^2 \le \frac{n(N_0 - n)}{N_0 - 1} \frac{1}{\binom{N_1}{N_0}} \sum_J \{d_2(F_0, F_{1J})^2 \colon |J| = N_0\}.$$

LEMMA 5. *For $\nu \ge 1$ let $\mathscr{X}_\nu$ be a finite population of size N_ν, F_ν the uniform distribution on $\mathscr{X}_\nu$, $X_1, \cdots, X_{n_\nu}$, a sample without replacement from $\mathscr{X}_\nu$, $\hat{F}_\nu$ the empirical* df *of the sample. If for some F, $d_2(F_\nu, F) \to 0$ as $\nu \to \infty$ and $n_\nu \to \infty$ then $d_2^2(\hat{F}_\nu, F) \to 0$ in probability.*

PROOF. If g is continuous and bounded

$$E \int g(x)\, d\hat{F}_\nu(x) = \int g(x)\, dF_\nu(x) \to \int g(x)\, dF(x),$$

$$\operatorname{Var}\left(\int g(x)\, d\hat{F}_\nu(x) \right) \to 0.$$

So,

$$\int g(x)\, d\hat{F}_\nu(x) \to \int g(x)\, dF(x) \tag{8}$$

in probability. Moreover,

$$\limsup_\nu E \int \phi(x, M)^2\, d\hat{F}_\nu(x) = \int \phi(x, M)^2\, dF(x)$$

by Lemma 8.3c) of Bickel and Freedman (1981). Since we can make $\int \phi(x, M)^2\, dF(x)$ small for M large we conclude that (8) holds for $g(x) = x^2$ also and the lemma follows. □

3. Proving the theorems in case (a).

PROOF OF THEOREM 1. Recall the variance weights w_i from Section 1. As is easily verified, $\hat{\tau}_a^2/\tau_a^2 = 1 + \xi - \zeta$, where

$$\xi = \textstyle\sum_{i=1}^p w_i^2 (n_i - 1)^{-1} \sum_{j=1}^{n_i} (Y_{ij}^2 - 1) \tag{9a}$$

$$\zeta = \textstyle\sum_{i=1}^p w_i^2 (n_i - 1)^{-1} (n_i Y_{i\cdot}^2 - 1). \tag{9b}$$

To prove the theorem, it is enough to show that ξ and ζ are both small. But $\xi = \xi_1 + \xi_2$, where

$$\xi_1 = \textstyle\sum_{i=1}^p (n_i - 1)^{-1} \sum_{j=1}^{n_i} [\bar{\phi}^2(w_i Y_{ij}, \varepsilon\sqrt{n_i}) - E\{\bar{\phi}^2(w_i Y_{ij}, \varepsilon\sqrt{n_i})\}] \tag{10a}$$

$$\xi_2 = \textstyle\sum_{i=1}^p (n_i - 1)^{-1} \sum_{j=1}^{n_i} [\phi^2(w_i Y_{ij}, \varepsilon\sqrt{n_i}) - E\{\phi^2(w_i Y_{ij}, \varepsilon\sqrt{n_i})\}]. \tag{10b}$$

Now

$$\begin{aligned}E(\xi_1^2) &= \operatorname{var}\xi_1 = \textstyle\sum_{i=1}^p (n_i-1)^{-2}\sum_{j=1}^{n_i}\operatorname{var}\{\bar\phi^2(w_iY_{ij},\varepsilon\sqrt{n_i})\}\\ &\le \textstyle\sum_{i=1}^p (n_i-1)^{-2} n_i E\{\bar\phi^4(w_iY_{ij},\varepsilon\sqrt{n_i})\}\\ &\le \varepsilon^2 \textstyle\sum_{i=1}^p (n_i-1)^{-2} n_i^2 E\{\bar\phi^2(w_iY_{ij},\varepsilon\sqrt{n_i})\}\\ &\le \varepsilon^2 \textstyle\sum_{i=1}^p (n_i-1)^{-2} n_i^2 w_i^2 E\{Y_{ij}^2\}\\ &\le 4\varepsilon^2 \textstyle\sum_{i=1}^p w_i^2 = 4\varepsilon^2.\end{aligned}$$

On the other hand, $E\{|\xi_2|\}\to 0$ for each $\varepsilon > 0$, by (4). This disposes of ξ.

The term ζ in (9b) can be decomposed according to whether $n_i > M$ or $n_i \le M$. Since

$$\textstyle\sum_i \{(n_i-1)^{-1}w_i^2\colon n_i \ge M+1\} \le M^{-1}$$

and $E\{n_iY_{i\cdot}^2\} = 1$, the strata i with $n_i \ge M+1$ are negligible. For the i with $n_i \le M$, $\zeta = \zeta_1 + \zeta_2$ where

$$\zeta_1 = \sum_i \frac{n_i}{n_i-1}\,[\bar\phi^2(w_iY_{i\cdot},\varepsilon\sqrt{n_i}) - E\{\bar\phi^2(w_iY_{i\cdot},\varepsilon\sqrt{n_i})\}] \tag{11a}$$

$$\zeta_2 = \sum_i \frac{n_i}{n_i-1}\,[\phi^2(w_iY_{i\cdot},\varepsilon\sqrt{n_i}) - E\{\phi^2(w_iY_{i\cdot},\varepsilon\sqrt{n_i})\}]. \tag{11b}$$

The sums need be extended only over i with $2 \le n_i \le M$. Now whatever n_i may be, as for ξ_1,

$$E\{\zeta_1^2\} \le 4\varepsilon^2 \tag{12}$$

is small. Next,

$$\begin{aligned}E\{|\zeta_2|\} &\le 2\sum_i \frac{n_i}{n_i-1}E\{\phi^2(w_iY_{i\cdot},\varepsilon\sqrt{n_i})\}\\ &\le 4M^2 \textstyle\sum_i E\{\phi^2(w_iY_{ij},\varepsilon\sqrt{n_i}/M)\}\end{aligned} \tag{13}$$

because $2 \le n_i \le M$; see Lemma 1. So ζ_2 is small too, by condition (4). □

PROOF OF THEOREM 2. The Lindeberg condition is applied, given $\mathscr{F}$. It is enough to check that for every $\varepsilon > 0$,

$$\tilde\tau_a^{-2}\textstyle\sum_{i=1}^p n_i^{-1}c_i^2 E\{\phi^2(X_{ij}^* - X_{i\cdot},\varepsilon n_i\tilde\tau_a|c_i|^{-1})\mid\mathscr{F}\}\to 0 \tag{14}$$

in probability, where $\tilde\tau_a^2 = \sum_{i=1}^p c_i^2(n_i-1)s_i^2/n_i^2$ is the conditional variance of $\hat\gamma^*$ given $\mathscr{F}$. For then, Theorem 1 can be applied to X_{ij}^*.

Since $n_i \ge 2$,

$$\tfrac{1}{2}\hat\tau_a \le \tilde\tau_a \le \hat\tau_a. \tag{15}$$

Thus $\hat{\tau}_a$ and hence τ_a may be substituted in (14) for $\tilde{\tau}_a$. So (14) reduces to

$$\tau_a^{-2} \sum_{i=1}^p c_i^2 n_i^{-2} \sum_{j=1}^{n_i} \phi^2(X_{ij} - X_{i\cdot}, \varepsilon n_i \tau_a | c_i |^{-1}) \to 0$$

in probability. This in turn reduces to

$$\sum_{i=1}^p n_i^{-1} \sum_{j=1}^{n_i} \phi^2[w_i(X_{ij} - X_{i\cdot})/\sigma_i, \varepsilon\sqrt{n_i}] \to 0 \tag{16}$$

in probability.

Now $(X_{ij} - X_{i\cdot})/\sigma_i = Y_{ij} - Y_{i\cdot}$. Use Lemma 1a) with $k = 2$ to see that (16) follows from (17) and (18):

$$\sum_{i=1}^p n_i^{-1} \sum_{j=1}^{n_i} \phi^2(w_i Y_{ij}, \tfrac{1}{4}\varepsilon\sqrt{n_i}) \to 0 \quad \text{in probability} \tag{17}$$

and

$$\sum_{i=1}^p \phi^2(w_i Y_{i\cdot}, \tfrac{1}{4}\varepsilon\sqrt{n_i}) \to 0 \quad \text{in probability.} \tag{18}$$

Clearly, (17) follows from (4). We bound the expected value of the left side of (18). Take first those i with $n_i \leq M$. In view of Lemma 1b), the sum over such i is bounded above by

$$M^2 \sum_i E\{\phi^2(w_i Y_{ij}, \tfrac{1}{4}\varepsilon\sqrt{n_i}/M)\}$$

which tends to zero by condition (4). Take next those i with $n_i > M$. The sum over such i is bounded above by

$$\sum_i E\{(w_i Y_{i\cdot})^2\} = \sum_i w_i^2 n_i^{-1} < M^{-1} \sum_i w_i^2 \leq M^{-1}$$

which is small for M large.

That $\tilde{\tau}_a^*/\tilde{\tau}_a \to 1$ follows from Theorem 1. □

REMARKS. (i) The Lindeberg-Feller theorem can be supplemented by direct bounds generalizing those of Berry-Esseen; see Petrov (1975, Theorem 3, page 111 or Theorem 8, page 118). These bounds may give estimates on the discrepancy between the bootstrap distribution and the true distribution.

(ii) The difference between the distribution of $(\hat{\gamma} - \gamma)/\tau_a$ and the bootstrap distribution of $(\hat{\gamma}^* - \hat{\gamma})/\tilde{\tau}_a$ can be estimated using the Mallows metric as in equation (2.2) of Bickel and Freedman (1981). The condition needed to push this through is stronger than (4).

(iii) The results can be extended in an obvious way to vector X_{ij}, and under further conditions to nonlinear statistics such as $\sum_{i=1}^p [g_i(X_{i\cdot}) - g_i(\mu_i)]$; this covers ratio estimates.

4. Proving the theorems in case (b)

PROOF OF THEOREM 3. The Lindeberg-Feller theorem does not apply to give us i) directly here, since the X_{ij} are dependent for fixed i; however, essentially the same ideas can be used. The proof we give is a bit complicated; an alternative but we believe no simpler approach is given by Dvoretzky (1971). Our argument is by cases, and the focus is on asymptotic normality. Without loss of generality, assume $\mu_i \equiv 0$, $c_i \equiv 1$. In outline, the argument is as follows.

CASE 1. There is only one stratum, and $n \le \frac{1}{2}N$; we drop the unnecessary stratum subscript i. Then ρ^2 is of order n, and asymptotic normality follows from Erdös-Renyi (1959). Also see Rosén (1967), Dvoretzky (1971).

CASE 2. There is only one stratum, and $n > \frac{1}{2}N$. Apply Case 1 to the "co-sample" consisting of the objects not in the sample.

CASE 3. The number of strata is bounded; no variance weight tends to zero. Case 1 or Case 2 applies to each stratum individually.

CASE 4. There are many strata, each of small variance weight; in each stratum, $n_i \le \frac{1}{2}N_i$. Then $\hat{\gamma}/\tau_b$ is the sum of p independent u.a.n. summands: var $\{X_{i\cdot}/\tau_b\} = v_i^2$ being uniformly small by assumption. We must verify the Lindeberg condition on $X_{i\cdot}/\tau_b$, and do so by an indirect argument. Let X'_{ij} be sampled with replacement from $\mathscr{X}_i$. And let

$$\hat{\gamma}' = \sum_{i=1}^{p} \frac{1}{n_i} \sum_{j=1}^{n_i} X'_{ij}.$$

Since $n_i \le \frac{1}{2}N_i$, the variance weights v_i^2 and w_i^2 are of the same order, as are the total variances τ_a^2 and τ_b^2. In particular, condition (6) implies (4). Thus, the Lindeberg condition holds for the individual summands in $\hat{\gamma}'/\tau_a$, viz., $X'_{ij}/n_i\tau_a$, and asymptotic normality of $\hat{\gamma}'$ follows. By the converse to Lindeberg's theorem, his condition holds for the stratum averages $(1/n_i)\sum_{j=1}^{n_i} X'_{ij}/\tau_a$. Hence, by Lemma 2, the condition holds for the stratum averages taken without replacement, viz., $(1/n_i)\sum_{j=1}^{n_i} X_{ij}/\tau_b$. Now a second application of the direct Lindeberg theorem gives asymptotic normality of $\hat{\gamma}$.

CASE 5. There are many strata, each of small variance weight; on each stratum, $n_i > \frac{1}{2}N_i$. Apply Case 4 to the co-samples.

CASE 6. There are many strata, each of small variance weight. Consider two groups of strata: in the first, $n_i \le \frac{1}{2}N_i$; in the second, $n_i > \frac{1}{2}N_i$. Case 4 applies to the first group, Case 5 to the second. (One of the two groups may be negligible.)

The general case. We combine cases 3 and 6. Let

$$J_k(\nu) = \left\{ i : v_i \ge \frac{1}{k} \right\}; \quad V_k(\nu) = \sum \{v_i^2 : i \in J_k(\nu)\}$$

where dependence on the hidden index is made explicit. Given any subsequence of $\{\nu\}$ we can extract a subsequence $\{\nu_r\}$ such that for all k, as $r \to \infty$, $V_k(\nu_r)$ tends to a finite limit V_k. If $V_k = 0$ for all k, there must be $k_r \to \infty$ such that $V_{k_r}(\nu_r) \to 0$. Hence, as $r \to \infty$,

$$\sum\{X_{i\cdot}/\tau_b: i \in J_{k_r}(\nu_r)\} \to 0 \quad \text{in probability.} \tag{19}$$

But, $\max\{v_i: i \notin J_{k_r}(\nu_r)\} \le 1/k_r \to 0$. So we can apply case 6 to get that

$$\sum\{X_{i\cdot}/\tau_b: i \notin J_{k_r}(\nu_r)\} \quad \text{is asymptotically} \quad N(0, 1). \tag{20}$$

Combining (19) and (20), we get

$$\sum X_{i\cdot}/\tau_b \quad \text{is asymptotically} \quad N(0, 1), \quad \text{as} \quad r \to \infty. \tag{21}$$

On the other hand, suppose $V_k > 0$ for some k. Since $J_k(\nu_r)$ has at most k^2 members, we can apply case 3 to see that for all k, as $r \to \infty$,

$$\sum\{X_{i\cdot}/\tau_b: i \in J_k(\nu_r)\} \quad \text{is asymptotically} \quad N(0, V_k).$$

By a standard argument, there are $k_r \to \infty$ such that

$$\sum\{X_{i\cdot}/\tau_b: i \in J_{k_r}(\nu_r)\} \quad \text{is asymptotically} \quad N(0, \sup_k \mathrm{V}_k). \tag{22}$$

Applying case 6 as above,

$$\sum\{X_{i\cdot}/\tau_b: i \notin J_{k_r}(\nu_r)\} \quad \text{is asymptotically} \quad N(0, 1 - \sup_k V_k). \tag{23}$$

Combining (22) and (23) we obtain (21) in this case also. Part (i) of the theorem follows by a standard compactness argument. The proof of (ii) follows the pattern of that of Theorem 1 and is omitted. □

PROOF OF THEOREM 5. We simplify the argument by supposing n_1 divides N_1 so we can use the naive bootstrap. (The general argument uses Lemma 4.) Moreover, without loss of generality let $\mu_1 = 0$, $\sigma_1 = 1$. Since $p = 1$ we want to compare the distribution of the standardized mean of a sample of size n_1 from the population $\mathscr{Y}_1$ and the distribution of the standardized mean of a sample of size n_1 from the population composed of N_1/n_1 copies of the standardized sample: $(X_{ij} - \hat{\mu}_1)/\hat{\sigma}_1$, $1 \le j \le n_1$, where $\hat{\mu}_1$ is the sample mean and $\hat{\sigma}_1$ is sample standard deviation. So by Lemma 3,

$$d_2^2\left\{\mathscr{L}\left(\frac{\hat{\gamma} - \gamma}{\tau_b}\right), \mathscr{L}\left(\frac{\hat{\gamma}^* - \hat{\gamma}}{\tilde{\tau}_b}\right) \,\middle|\, X_{1j}, 1 \le j \le n_1\right\} \le d_2^2\{F_\nu, \hat{F}_\nu, \hat{F}_\nu(\hat{\sigma}_1 x + \hat{\mu}_1)\}.$$

By Lemma 5, $d_2^2(F_\nu, \hat{F}_\nu)$, $\hat{\mu}$, and $\hat{\sigma}_1 - 1$ all tend in probability to 0 as $\nu \to \infty$. A truncation argument of the type we have already used shows that $\hat{\tau}_b/\tau_b$ and $\hat{\tau}_b^*/\hat{\tau}_b$ both tend in probability to 1. The theorem follows. □

REFERENCES

BABU, G. J. and SINGH, K. (1983). Inference on means using the bootstrap. *Ann. Statist.* **11** 999–1003.

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217.

BICKEL, P. J. and KRIEGER, A. (1983). Using the bootstrap to set confidence bands for a distribution function: Monte Carlo and some theory. In preparation.

CHAO, M. T. and LO, S. H. (1983). A bootstrap method for finite population. Preprint.

DVORETZKY, A. (1971). Asymptotic normality for sums of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Probab.* II 513. (Ed: Le Cam, Neyman, Scott). Univ. of Calif. Press, Berkeley.

ERDÖS, P. and RENYI, A. (1959). On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hung. Acad. Sci.* **4** 49–61.

GROSS, S. (1980). Median estimation in sample surveys. Paper presented at 1980 Amer. Statist. Assoc. meeting.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

KREWSKI, D. and RAO, J. N. K. (1981). Inference from stratified samples: properties of linearization, jackknife, and balanced repeated replication. *Ann. Statist.* **9** 1010–1019.

MALLOWS, C. (1972). A note on asymptotic joint normality. *Ann. Math. Statist.* **43** 508–515.

PETROV, V. (1975). *Sums of Independent Random Variables.* Springer, Berlin.

ROSÉN, B. (1967). On the central limit theorem for sums of dependent random variables. *Z. Wahrsch. verw. Gebiete* **7** 48–82.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

Richardson Extrapolation and the Bootstrap

PETER J. BICKEL and JOSEPH A. YAHAV*

Simulation methods [particularly Efron's (1979) bootstrap] are being applied more and more frequently in statistical inference. Given data $(X_1, \ldots, X_n)$ distributed according to P, which belongs to a hypothesized model $\mathbf{P}$, the basic goal is to estimate the distribution $\mathbf{L}_P$ of a function $T_n(X_1, \ldots, X_n, P)$. The bootstrap presupposes the existence of an estimate $\hat{P}(X_1, \ldots, X_n)$ and involves estimating $\mathbf{L}_P$ by the distribution $\mathbf{L}_n^*$ of $T_n(X_1^*, \ldots, X_n^*, \hat{P})$, where $(X_1^*, \ldots, X_n^*)$ is distributed according to $\hat{P}$. The method is of particular interest when L_n^*, though known in principle, can realistically only be computed by simulation. Such computation can be expensive if n is large and T_n is complex (e.g., see the multivariate goodness-of-fit tests of Beran and Millar 1986). Even when bootstrap application to a single data set is not excessively expensive, Monte Carlo studies of the bootstrap are another matter. We propose a method based on the classical ideas of Richardson extrapolation for reducing the computational cost inherent in bootstrap simulations and Monte Carlo studies of the bootstrap, by performing simulations for statistics based on two smaller sample sizes. We study theoretically which ratio of the two small sample sizes is apt to give best results. We show how our method works for approximating the χ^2, t, and smoothed binomial distributions, and for setting bootstrap percentile confidence intervals for the variance of a normal distribution with a mean of 0.

KEY WORDS: Cost of computation; Edgeworth expansion; Approximation.

1. INTRODUCTION

Let L_n^* be the bootstrap distribution of $T_n(X_1, \ldots, X_n, P)$. With knowledge of particular features of L_n^*, various devices such as importance sampling can reduce the number r of Monte Carlo replications needed to compute (or rather estimate) L_n^* closely. The total computation cost for a simulation is proportional to $c(n)r$, where $c(n)$, the cost of computing T_n, usually rises at least linearly with n (and often faster). In this article we explore a way of reducing $c(n)$ rather than r. Suppose that T_n is univariate, and let F_n^* be the distribution function of L_n^*. For most T_n of interest, it is either known or plausibly conjectured that F_n^* tends to a limit A_0 in probability

$$F_n^*(x) = A_0(x) + o_p(1), \qquad (1.1)$$

for all x and often uniformly in x as well. Examples include the usual pivots for parameters $\theta(F)$ when $X_1, \ldots, X_n$ are iid F and $\hat{P} \leftrightarrow \hat{F}$ is the empirical distribution. Thus if $T_n = \sqrt{n}(\theta(\hat{F}) - \theta(F))$, then $A_0 = \mathbf{N}(0, \sigma^2(F))$, under mild conditions; if $T_n = \sqrt{n}\,[(\theta(\hat{F}) - \theta(F))/\sigma(\hat{F})]$, then $A_0 = \mathbf{N}(0, 1)$. The value A_0 can also be known to exist but not be readily computable. For example, let $T_n = \sqrt{n}\ \sup_x|\hat{F}(x) - F(x)|$, with F possibly discrete (see Bickel and Freedman 1981). Furthermore, an asymptotic expansion in powers of $n^{-1/2}$ is known to be true in some cases and reasonably conjectured in many others. That is,

$$F_n^*(x) = A_0(x) + \sum_{j=1}^{k} n^{-j/2}A_j(x) + O_P(n^{[-(k+1)]/2}). \qquad (1.2)$$

The most important special cases arise when A_0 is normal and the expansion (1.2) is of Edgeworth type. Such expansions appear in the bootstrap context in works by Singh (1981), Bickel and Freedman (1981), and Abramovitch and Singh (1985), among others. Expansions for the distributions F_n of $T_n(X_1, \ldots, X_n)$ under fixed F have been studied extensively (e.g., see Bhattacharya and Ranga Rao 1976).

In this context, we propose to calculate $F_{n_1}, \ldots, F_{n_{k+1}}$, where

$$n_1 + \cdots + n_{k+1} = b \ll n. \qquad (1.3)$$

We use the F_{n_j} to approximate F_n. This procedure is classically used in numerical analysis (where it is called Richardson extrapolation) to approximate F_∞. Our application of these ideas differs, in that

1. We are interested in F_n, not F_∞.
2. F_∞ is sometimes known, as in the Edgeworth case, and can be used to improve the approximation.
3. We are interested in the design problem of selecting the n_j, subject to the budget constraint (1.3).

Using our method in the bootstrap context involves simply putting $*$ on the F_{n_j} and F_n. In Section 2 we develop the method in detail and give explicit solutions to three formulations of the design problem for $k = 1$. Finally, in Section 3 we test the method on approximations of known F_n, as well as some bootstrap examples. The results are very encouraging.

2. EXTRAPOLATION

Throughout this section IK refers to Isaacson and Keller (1966). Write $t = n^{-1/2}$ $(0 < t \leq 1)$. Given a sequence of distribution functions $F_n \triangleq G_t$, write

$$G_t = P_t + \Delta_t, \qquad P_t = A_0 + \sum_{j=1}^{k} t^j A_j. \qquad (2.1)$$

The argument in the functions G_t and A_j plays no role in our discussion and is omitted. We calculate $G_{t_0}, \ldots,$

* Peter J. Bickel is Professor, Department of Statistics, University of California, Berkeley, CA 94720. Joseph A. Yahav is Professor, Department of Statistics, Hebrew University, Jerusalem, Israel. This work was partially supported by Office of Naval Research Contract N00014-80-C0163. The authors thank Persi Diaconis for a reference to Kuipers and Niederreiter (1974), which they used to obtain a considerable simplification of the original proof of the theorem in the Appendix, and Adele Cutler, for the programming of the simulations and other calculations in Section 3.

Journal of the American Statistical Association
June 1988, Vol. 83, No. 402, Theory and Methods

G_{t_k} $(t < t_0 < \cdots < t_k)$. If $\Delta_t = 0$ for $t, t_0, \ldots, t_k$ we obtain G_t perfectly from the G_{t_j} by using the Lagrange interpolating polynomial (IK, p. 188):

$$\hat{G}_t = \sum_{j=0}^{k} G_{t_j}\phi_{k,j}(t), \qquad \phi_{k,j}(t) = \prod_{i \neq j} [(t - t_i)/(t_j - t_i)]. \tag{2.2}$$

In particular for the only case we study in detail, $k = 1$,

$$\hat{G}_t = (t_1 - t_0)^{-1}[(t_1 - t)G_{t_0} + (t - t_0)G_{t_1}]. \tag{2.3}$$

We consider three classes for Δ, depending on a parameter M:

1. $\mathbf{D}_1 = \{\Delta: d^{k+1}\Delta_t/dt^{k+1}$ exists and $\sup_t|(d^{k+1}\Delta_t)/dt^{k+1}| \leq M\}$. Since Δ is only defined at the points $n^{-1/2}$ $(n = 1, 2, \ldots)$ we interpret $\Delta \in \mathbf{D}_1$ as applying to some smooth function agreeing with Δ at all points $n^{-1/2}$. The other two classes make no smoothness assumptions on Δ.
2. $\mathbf{D}_2 = \{\Delta: \sup_t t^{-(k+1)}|\Delta_t| \leq M\}$.
3. $\mathbf{D}_3 = \{\Delta: 0 \leq t^{-(k+1)}\Delta_t \leq M$ for all $t > 0$, or $-M \leq t^{-(k+1)}\Delta_t \leq 0$ for all $t > 0\}$.

For fixed $t, t_0, \ldots, t_k$ we define the error of approximation by

$$E_i(t, t_0, \ldots, t_k) = \sup\{|\hat{G}_t - G_t|: \Delta \in D_i\}, \quad 1 \leq i \leq 3.$$

We want to minimize E_i, subject to a fixed budget b, where

$$\sum_{j=0}^{k} t_j^{-2} = b. \tag{2.4}$$

If t_j satisfy (2.4) and $b \to \infty$, then $t_0 \to 0$.

We claim that

$$E_1 \sim \frac{M}{(k+1)!}\prod_{j=0}^{k}(t_j - t), \tag{2.5}$$

$$E_2 \sim M\left\{\sum_{j=0}^{k} |\phi_{k,j}(t)|t_j^{k+1} + t^{k+1}\right\}, \tag{2.6}$$

and

$$E_3 \sim M\left\{\left[\sum_{j=0}^{k} [\phi_{k,j}(t)]_+ t_j^{k+1}\right] \vee \left[\sum_{j=0}^{k} [\phi_{k,j}(t)]_- t_j^{k+1}\right] + t^{k+1}\right\}, \tag{2.7}$$

where $a_+ = a \vee 0$ and $a_- = -(a \wedge 0)$. To check (2.5), apply theorem 1 of IK (p. 90), which has

$$G_t - \hat{G}_t = [(k+1)!]^{-1}\prod_{i=0}^{k}(t - t_i)\frac{d^{k+1}G_t}{dt^{k+1}}(\xi), \tag{2.8}$$

where $t < \xi < t_k$. Note that $(d^{k+1}/dt^{k+1})P_t = 0$. To check (2.6) and (2.7), note that interpolation is linear, so $\hat{G}_t = \hat{P}_t + \hat{\Delta}_t$. Since $P_t = \hat{P}_t$, we have $G_t - \hat{G}_t = \Delta_t - \hat{\Delta}_t$; (2.6) and (2.7) follow from (2.2). From (2.5), E_1 is minimized subject to (2.3) as $b \to 0$ by

$$t_0 = \cdots = t_k = \sqrt{(k+1)/b}. \tag{2.9}$$

must be distinct. Nevertheless, if the error term Δ is sufficiently smooth, the n_j should be chosen as nearly equal to each other as possible.

This procedure is analogous to that of the "leave-one-out" jackknife process. This conclusion is clearly valid not just under (2.4), but under any reasonable symmetric-side condition on $t_0, \ldots, t_k$. If we suppose that $t = o(t_0)$, that is, the budget is much smaller than n, we can simplify (2.6) to

$$E_2 \sim M\left(\prod_{j=0}^{k} t_j\right)\sum_{j=0}^{k} t_j^k\left[\prod_{i<j}(t_j - t_i)\prod_{i>j}(t_i - t_j)\right]^{-1} \tag{2.10}$$

and (2.7) to

$$E_3 \sim M\left(\prod_{j=0}^{k} t_j\right)\sum_{j=0}^{k} t_j^k \times \min\left\{\left[\prod_{i\neq j}(t_j - t_i)\right]_+, \left[\prod_{i \neq j}(t_j - t_i)\right]_-\right\}. \tag{2.11}$$

Evidently, (2.10) is minimized asymptotically by $t_j^{-2} = \lambda_j^2 b$, where $\lambda_j > 0$,

$$\sum_{j=0}^{k}\lambda_j^2 = 1, \tag{2.12}$$

and $\lambda_0, \ldots, \lambda_k$ minimize

$$\left(\prod_{j=0}^{k}\lambda_j\right)^{-1}\sum_{j=0}^{k}\left[\lambda_j\prod_{i<j}(\lambda_i - \lambda_j)\prod_{i>j}(\lambda_j - \lambda_i)\right]^{-1}, \tag{2.13}$$

subject to (2.12). In principle, this minimization can be carried out for any k. The explicit solutions for the cases we are primarily concerned with, E_2 and E_3 for $k = 1$, are as follows (if we ignore the restriction that the $\lambda_j^2 b$ are integers): For E_2,

$$\lambda_0^2 = 1 - \lambda_1^2 = .89, \tag{2.14}$$

or more specifically $\lambda_0 = \cos[\frac{1}{2}(\sin^{-1}(1/\omega_0))]$, where $\omega_0 = (1 + \sqrt{5})/2 = 1.6180$ is the unique positive root of $\omega^3 - 2\omega - 1 = 0$. To see this note that for $k = 1$, (2.13) is simply $(\lambda_0\lambda_1)^{-1}(\lambda_0 - \lambda_1)^{-1}(\lambda_0 + \lambda_1)$. Substitute $\lambda_0 = \cos\theta$ to get the objective,

$$2(1 + \sin 2\theta)(\cos 2\theta \sin 2\theta)^{-1},$$

and then substitute $\sin 2\theta = (1 - \nu^2)^{1/2} = 1/\omega$. Similarly, for $k = 1$, $E_3 \sim M[t_0 t_1^2/(t_1 - t_0)]$; a similar minimization gives

$$\lambda_0^2 = \tfrac{1}{2}(1 + (1/\sqrt{2})) = .85. \tag{2.15}$$

In all of these cases, $E_j = O(b^{-(k+1)/2})$.

We check our approach in the following examples of $\{F_n\}$, belonging to $\mathbf{D}_1$ and $\mathbf{D}_3$, respectively.

Example 1: The Gamma Family. Let F_n be the distribution of $(S_n - n)(2n)^{-1/2}$, where S_n has the χ_n^2 distribution. Evidently, we can define G_t for $t > 0$ with

$$G_t(x) = \Gamma^{-1}(\nu^{-1})\lambda^{\nu} \int_0^{x_t} e^{-\lambda s} s^{\nu-1}\, ds \tag{2.16}$$

where $x_t = x\nu^{1/2} + (\nu/\lambda)$, $\nu = 2t^{-2}$, and $\lambda = \frac{1}{2}$. Using standard Stirling expansions for Γ and its derivatives, it is easy to show that

$$G_t(x) = \frac{e^{-\nu}\nu^{\nu-1/2}}{\Gamma(\nu)} \int_{-\sqrt{\nu}}^{x} \times [(\exp(-u\nu^{-1/2}))(1 + u\nu^{-1/2})]^{\nu}(1 + u\nu^{-1/2})^{-1}\, du,$$

that $A_0 = \Phi$, the standard normal distribution, and that G_t has bounded derivatives of all orders in t. Thus $\Delta \in \mathbf{D}_1$ for all k. Evidently, taking $\lambda = \frac{1}{2}$ plays no role, and this observation applies to the standardized gamma family in general.

Example 2: The Binomial Distribution With Continuity Correction. Let F_n be the distribution of $(S_n - np)/(npq)^{1/2}$ convoluted with the uniform distribution on

$$[-1/(2\sqrt{npq}),\ 1/(2\sqrt{npq})],$$

where S_n has a binomial (n, p) distribution $q = 1 - p$ $(0 < p < 1)$. It is well known that F_n is of the form $F_n(x) = \Phi(x) + n^{-1/2}A_1(x) + O(n^{-1})$ (e.g., see Feller 1971, p. 540). But if we analyze the remainder term further, by theorem 23.1 of Bhattacharya and Ranga Rao (1971, p. 238) it is of the form

$$F_n(x) - \Phi(x) - n^{-1/2}A_1(x) = n^{-1}\left[\int_{-1/2}^{1/2} uS_1(np + x\sigma\sqrt{n} - u)\, du\right] \times P(x, \sigma) + o(n^{-1}), \tag{2.17}$$

where $\sigma = (pq)^{1/2}$, $S_1(t) = t - \frac{1}{2}$ $(0 < t < 1)$, and $S_1(t + 1) = S_1(t)$. Check that

$$\int_{-1/2}^{1/2} uS_1(\nu - u)\, du = -\frac{S_1^2}{2}\left(x + \frac{1}{2}\right). \tag{2.18}$$

Unless $x = 0$ and p is rational, the sequence $S_1(np + \sqrt{n}\sigma x + \frac{1}{2})$ is uniformly distributed modulo 1; that is, $\#\{h : S_1(np + \sqrt{n}\sigma x + \frac{1}{2}) \le t,\ n \le N\}/N \to t + \frac{1}{2}$ as $N \to \infty$ if $(-\frac{1}{2} < t < \frac{1}{2})$. A proof is given in the Appendix.

Thus as $n \to \infty$ the coefficient of n^{-1} in (2.17) ranges over an interval $[0, \frac{1}{8}]$ or $[-\frac{1}{8}, 0]$, and comes arbitrarily close to all values in the interval. Hence, $\{F_n\}$ belongs to $\mathbf{D}_3$ for $k = 1$.

Notes. In many examples (including the two we have discussed) A_0 is known. Then, if (2.1) holds for $k = r + 1$, we can improve our estimate using only k sample sizes and still have an error $O(b^{-(r+1)/2})$. We define $Q_t = (G_t - A_0)/t$ and use the estimate $G_t^* = A_0 + t\hat{Q}_t$, where $\hat{Q}_t$ is defined by (2.2), with $k = r$. In particular, for $r = 1$ the allocations (2.9), (2.14), and (2.15) give errors $O(b^{-3/2})$. In the next section we study this approximation by simulation as well.

In some cases such as F_n, the t distribution with n degrees of freedom, the series is in powers of n^{-1}. In this case it is easy to obtain the optimal choice of t_0/t_1 for $\mathbf{D}_2$ and $\mathbf{D}_3$, that is, for (2.4) replaced by $t_0^{-1} + t_1^{-1} = b$. We find for $\mathbf{D}_2$

$$n_j = \rho_j b, \quad \rho_1 = 1 - \rho_0, \quad \rho_0 = .5(1 + \sqrt{3}) = .79, \tag{2.19}$$

and for $\mathbf{D}_3$

$$\rho_0 = .75. \tag{2.20}$$

If (as is usually the case in applications) the A_j and t are unknown, it would seem safer to use the approximation for $t = n^{-1/2}$.

An undesirable feature of our approach is that no a posteriori estimate of the error actually incurred is available. If t_1 is small and $\Delta \in \mathbf{D}_1$, we can get an estimate by increasing our budget. We add $\bar{t}^{-2} \ne t_j^{-2}$ $(j = 0, 1)$ units and calculate $G_{\bar{t}}$. Now, by (2.8),

$$G_u - \hat{G}_u = \tfrac{1}{2}(d^2\Delta/dt^2)(\xi)(t_1 - t_0)^{-1}(u - t_0)(u - t_1), \tag{2.21}$$

where $t < \xi < t_1$ for any $t \le u \le t_1$. If t_1 is small we expect the coefficient $d^2\Delta/dt^2$ in (2.21) to be stable, so we obtain

$$|G_t - \hat{G}_t| \propto |(t - t_0)(t - t_1)(s - t_0)^{-1}(s - t_1)^{-1}| \times |G_{\bar{t}} - \hat{G}_{\bar{t}}|. \tag{2.22}$$

If $\Delta \in \mathbf{D}_2$ or $\mathbf{D}_3$, no realistic estimate of the error presents itself. Suppose, however (as may be seen in Ex. 2), that if $0 < \lambda_1 < \cdots < \lambda_k < 1$, $a_1, \ldots, a_k$ are real, $n \to \infty$, and $s_j = [\lambda_j n]^{-1/2}$, then

$$\#(\Delta_{s_j} \le s_j^2 a_j : 1 \le j \le k)/n \to \prod_{j=1}^{k} G(a_j). \tag{2.23}$$

That is, $s_1^{-2}\Delta_{s_1}, \ldots, s_k^{-2}\Delta_{s_k}$ are asymptotically independently distributed with common distribution G. This is, of course, a poor approximation if λ_j and λ_{j+1} are too close and we cannot use (2.23) for design. But if we increase our budget we can calculate G at $l \geqslant 3$ points $t_0, t_1, \ldots, t_l$, with $1 \ge 2$. If we assume (2.1), it is natural to consider the estimate

$$\hat{G}_t^l = \hat{A}_0^l + t\hat{A}_1^l, \tag{2.24}$$

where $\hat{A}_0^l$ and $\hat{A}_1^l$ are the weighted fixed least squares estimates of A_0 and A_1,

$$\hat{A}_1^l = \sum_{i=0}^{l} (t_i - \bar{t})G_{t_i}\sigma_i^{-2} \Big/ \sum_{i=0}^{l} (t_i - \bar{t})^2\sigma_i^{-2}, \tag{2.25}$$

and

$$\hat{A}_0^l = \sum_{i=0}^{l} G_{t_i}\frac{\sigma_i^{-2}}{W} - \hat{A}_1^l\bar{t}, \tag{2.26}$$

Table 1. Richardson Extrapolation for χ^2_n

n_0, n_1	Percentiles 10	90	95	99
$n = 50$				
True values	−1.2311	1.3167	1.7505	2.6154
Fisher approximation	−1.1995	1.3637	1.7802	2.5969
	(.0317)	(.0470)	(.0297)	(−.0185)
	$n_0 + n_1 = 15$			
1, 14	−1.2557	1.3572	1.7995	2.6686
	(−.0246)	(.0404)	(.0490)	(.0532)
3, 12	−1.2528	1.3417	1.7796	2.6446
	(−.0217)	(.0250)	(.0291)	(.0292)
4, 11	−1.2511	1.3391	1.7764	2.6410
	(−.0199)	(.0224)	(.0259)	(.0256)
6, 9	−1.2493	1.3367	1.7735	2.6377
	(−.0182)	(.0200)	(.0230)	(.0223)
	$n_0 + n_1 = 20$			
2, 18	−1.2481	1.3374	1.7747	2.6400
	(−.0169)	(.0207)	(.0242)	(.0246)
4, 16	−1.2448	1.3319	1.7680	2.6324
	(−.0137)	(.0152)	(.0175)	(.0170)
5, 15	−1.2439	1.3307	1.7665	2.6307
	(−.0127)	(.0139)	(.0160)	(.0153)
8, 12	−1.2424	1.3289	1.7644	2.6258
	(−.0113)	(.0122)	(.0139)	(.0131)
$n = 100$				
True values	−1.2475	1.3080	1.7212	2.5319
Fisher approximation	−1.2235	1.3397	1.7406	2.5176
	(.0239)	(.0317)	(.0193)	(−.0143)
	$n_0 + n_1 = 20$			
2, 18	−1.2740	1.3399	1.7585	2.5694
	(−.0265)	(.0319)	(.0373)	(.0374)
4, 16	−1.2687	1.3314	1.7481	2.5576
	(−.0212)	(.0234)	(.0268)	(.0257)
5, 15	−1.2671	1.3294	1.7457	2.5550
	(−.0197)	(.0214)	(.0245)	(.0231)
8, 12	−1.2649	1.3267	1.7424	2.5515
	(−.0174)	(.0187)	(.0212)	(.0196)
	$n_0 + n_1 = 30$			
3, 27	−1.2628	1.3252	1.7410	2.5508
	(−.0154)	(.0172)	(.0197)	(.0189)
6, 24	−1.2592	1.3205	1.7354	2.5448
	(−.0117)	(.0125)	(.0142)	(.0129)
7, 23	−1.2585	1.3198	1.7345	2.5439
	(−.0110)	(.0117)	(.0133)	(.0120)
12, 18	−1.2569	1.3180	1.7324	2.5418
	(−.0095)	(.0100)	(.0112)	(.0099)

where $\sigma_i = t_i^2$, $W = \sum_{i=0}^{l} \sigma_i^{-2}$, and $\bar{t} = \sum_{i=0}^{l} t_i\sigma_i^{-2}/W$. The error, $G_t - \hat{G}_t^l$, can be estimated by

$$\left\{ W^{-1} + t^2[\Sigma(t_i - \bar{t})^2\sigma_i^{-2}]^{-1} \times \sum_{i=0}^{l} (G_{t_i} - \hat{A}_0^l - \hat{A}_1^l t_i)^2 \sigma_i^{-2} \right\}^{1/2}. \quad (2.27)$$

The range of validity of the approximations (2.22) and (2.27) needs to be investigated by simulation.

3. COMPUTATION AND SIMULATION

In this section we study the actual performance of the approximations in the Section 2 examples. We also study the performance of the approximation for the Student-t distribution, where the expansion is in powers of $1/n$. Finally, we provide the results of a bootstrap simulation, where we compare the operating characteristics of confidence bounds based on a Richardson extrapolation approximation with those based on a full bootstrap.

χ^2_n Approximation. We computed the Richardson extrapolation for $[\chi^2_n(\alpha) - n]/(2n)^{1/2}$ (α = 10%, 90%, 95%, 99%), where $\chi^2_n(\alpha)$ is the αth percentile of the χ^2_n distribution, and compared it with the Fisher square-root approximation applied to the quantiles:

$$[\chi^2_n(\alpha) - n]/\sqrt{2n} \propto Z(\alpha) + [Z^2(\alpha)]/2\sqrt{2n},$$

where $Z(\alpha)$ is the standard normal α percentile. We used $n = 50, 100$, $b = 15, 20, 30$, and $1 - \lambda = n_0/b = .1, .2, .25, .40$, where $n_0 < n_1$ and $n_0 + n_1 = b$. Note the following:

1. The approximation improves as b and n increase.
2. The allocation $\lambda = .6$ is best, as expected.
3. For $n_0 + n_1 = 15$, 20, and all λ, the Richardson extrapolation is essentially as good as Fisher's approximation for the .9 and .1 percentiles, and still gives the same two significant figures as Fisher's for the .95 and .99 percentiles.
4. For $n_0 + n_1 = 30$ it is better in all cases save one, where the results are virtually equivalent. The $\lambda = .6$ allocation seems to give nearly three significant figures (see Table 1).

Table 2. Richardson Extrapolation for χ^2_{50} Knowing the Limit

n_0, n_1	Percentiles 10	90	95	99
$n = 50$				
True values	−1.2311	1.3167	1.7505	2.6154
Fisher approximation	−1.1995	1.3637	1.7802	2.5969
	(.0317)	(.0470)	(.0297)	(−.0185)
	$n_0 + n_1 = 15$			
1, 14	−1.2289	1.3165	1.7510	2.6178
	(.0022)	(−.0002)	(.0005)	(.0024)
3, 12	−1.2306	1.3168	1.7510	2.6172
	(.0006)	(.0001)	(.0005)	(.0018)
4, 11	−1.2307	1.3168	1.7509	2.6171
	(.0004)	(.0001)	(.0005)	(.0017)
6, 9	−1.2308	1.3168	1.7509	2.6169
	(.0003)	(.0001)	(.0004)	(.0016)
	$n_0 + n_1 = 20$			
2, 18	−1.2305	1.3167	1.7509	2.6168
	(.0006)	(.0000)	(.0004)	(.0015)
4, 16	−1.2309	1.3168	1.7508	2.6166
	(.0002)	(.0001)	(.0003)	(.0012)
5, 15	−1.2309	1.3168	1.7508	2.6165
	(.0002)	(.0001)	(.0003)	(.0011)
8, 12	−1.2310	1.3168	1.7508	2.6164
	(.0001)	(.0001)	(.0003)	(.0010)
	$n_0 + n_1 = 30$			
3, 27	−1.2310	1.3167	1.7507	2.6161
	(.0002)	(.0000)	(.0002)	(.0007)
6, 24	−1.2311	1.3167	1.7506	2.6159
	(.0001)	(.0000)	(.0002)	(.0005)
7, 23	−1.2311	1.3167	1.7506	2.6159
	(.0001)	(.0000)	(.0001)	(.0005)
12, 18	−1.2311	1.3167	1.7506	2.6159
	(.0000)	(.0000)	(.0001)	(.0005)

Table 3. Richardson Extrapolation for the t Distribution

n_0, n_1	Percentiles 10	90	95	99
$n = 50$				
True values	−1.2987	1.2987	1.6759	2.4033
Normal approximation	−1.2816	1.2816	1.6449	2.3263
	(.0171)	(−.0171)	(−.0310)	(−.0770)
	$n_0 + n_1 = 15$			
3, 12	−1.2849	1.2849	1.6376	2.2099
	(.0138)	(−.0138)	(−.0383)	(−.1934)
4, 11	−1.2878	1.2878	1.6462	2.2595
	(.0110)	(−.0110)	(−.0298)	(−.1438)
6, 9	−1.2900	1.2900	1.6526	2.2947
	(.0087)	(−.0087)	(−.0233)	(−.1086)
	$n_0 + n_1 = 20$			
4, 16	−1.2922	1.2922	1.6584	2.3198
	(.0065)	(−.0065)	(−.0175)	(−.0835)
5, 15	−1.2933	1.2933	1.6614	2.3357
	(.0055)	(−.0055)	(−.0146)	(−.0676)
8, 12	−1.2945	1.2945	1.6649	2.3535
	(.0042)	(−.0042)	(−.0111)	(−.0498)
$n = 100$				
True values	−1.2901	1.2901	1.6602	2.3642
Normal approximation	−1.2816	1.2816	1.6449	2.3263
	(.0085)	(−.0085)	(−.0153)	(−.0379)
	$n_0 + n_1 = 20$			
4, 16	−1.2818	1.2818	1.6378	2.2577
	(.0083)	(−.0083)	(−.0224)	(−.1065)
5, 15	−1.2831	1.2831	1.6417	2.2785
	(.0070)	(−.0070)	(−.0185)	(−.0857)
8, 12	−1.2848	1.2848	1.6463	2.3018
	(.0053)	(−.0053)	(−.0139)	(−.0624)
	$n_0 + n_1 = 30$			
6, 24	−1.2869	1.2869	1.6520	2.3274
	(.0031)	(−.0031)	(−.0082)	(−.0368)
7, 23	−1.2873	1.2873	1.6530	2.3321
	(.0028)	(−.0028)	(−.0072)	(−.0321)
12, 18	−1.2880	1.2880	1.6550	2.3414
	(.0020)	(−.0020)	(−.0053)	(−.0228)

In Table 2 we exhibit the Richardson extrapolation results for the χ^2_n distribution, using the knowledge of the limit as $n \to \infty$ (see Sec. 2).That is, we use the expansion

$$[\chi^2_n(\alpha) - n]/\sqrt{2n} = Z(\alpha) + A_1(1/\sqrt{n}) + A_2\frac{1}{n} + o_P\left(\frac{1}{n}\right)$$

or

$$\sqrt{n}\,\{[\chi^2_n(\alpha) - n]/\sqrt{2n} - Z(\alpha)\} = A_1 + A_2(1/\sqrt{n}) + o_P(1/\sqrt{n}),$$

where $Z(\alpha)$ is the α percentile of the standard normal. A_1 and A_2 are estimated using $\chi^2_{n_0}$ and $\chi^2_{n_1}$. The results are extremely good for both $n = 50$ and $n = 100$ (omitted here). The extrapolation, even for $n_0 + n_1 = 15$ and $\lambda = .9$, gives three significant figures for all percentiles. For $n_0 + n_1 = 30$, it often gives five significant figures.

The Student-t distribution has an expansion in powers of $1/n$. The Richardson extrapolation (2.3) with $1/\sqrt{n}$ gave no improvement over the ordinary normal approximation, as expected. In Table 3 we present the Richardson extrapolation to the t distribution and compare these results with the normal approximation. We looked at the same values of n, b, λ, and α for approximation to $t_n(\alpha)$, the αth percentile of the t distribution with n degrees of freedom. For $\lambda = .6$ and $b = 30$, the approximation is valid to 3 significant figures for $n = 100$ in all but one case, and improves on the normal approximation.

Tables 4 and 5 give the Richardson extrapolation for the continuity-corrected binomial distribution. That is, we define

$$B_n(s) = \sum_{k=0}^{[s]} \binom{n}{k} p^k(1-p)^{n-k} + (s - [s])\binom{n}{[s]+1} p^{[s]+1}(1-p)^{n-1-[s]}$$

Table 4. Richardson Extrapolation for the Binomial Distribution With p = .2

n_0, n_1	Percentiles 10	90	95	99
$n = 50$				
True values	−1.2591	1.3125	1.7177	2.4900
Normal approximation	−1.2816	1.2816	1.6449	2.3263
	(−.0225)	(−.0309)	(−.0728)	(−.1637)
	$n_0 + n_1 = 15$			
1, 14	−1.2689	1.2591	1.6071	2.4969
	(−.0097)	(−.0533)	(−.1106)	(.0068)
3, 12	−1.2392	1.3702	1.6743	2.6561
	(.0199)	(.0577)	(−.0434)	(.1661)
4, 11	−1.1692	1.2861	1.5821	2.3349
	(.0900)	(−.0264)	(−.1356)	(−.1551)
6, 9	−1.1679	1.4169	1.6882	2.5377
	(.0913)	(.1044)	(−.0295)	(.0477)
	$n_0 + n_1 = 20$			
2, 18	−1.2182	1.3060	1.6984	2.4728
	(.0409)	(−.0065)	(−.0193)	(−.0172)
4, 16	−1.2751	1.2595	1.7357	2.4304
	(−.0160)	(−.0530)	(.0180)	(−.0597)
5, 15	−1.2724	1.3224	1.7362	2.5814
	(−.0133)	(.0099)	(.0185)	(.0914)
8, 12	−1.1082	1.2538	1.8704	2.8400
	(.1509)	(−.0587)	(.1527)	(.3500)
$n = 100$				
True values	−1.2733	1.3036	1.6922	2.4351
Normal approximation	−1.2816	1.2816	1.6449	2.3263
	(−.0083)	(−.0220)	(−.0473)	(−.1088)
	$n_0 + n_1 = 20$			
2, 18	−1.2111	1.2835	1.6845	2.4144
	(.0622)	(−.0202)	(−.0078)	(−.0207)
4, 16	−1.2699	1.2280	1.7121	2.3686
	(.0033)	(−.0757)	(.0199)	(−.0665)
5, 15	−1.2638	1.3054	1.7208	2.5622
	(.0094)	(.0018)	(.0286)	(.1271)
8, 12	−1.0651	1.2220	1.8840	2.8864
	(.2082)	(−.0816)	(.1918)	(.4513)
	$n_0 + n_1 = 30$			
3, 27	−1.2644	1.3313	1.6692	2.4688
	(.0089)	(.0277)	(−.0231)	(.0337)
6, 24	−1.2233	1.3226	1.6628	2.4172
	(.0500)	(.0190)	(−.0294)	(−.0179)
7, 23	−1.2386	1.2849	1.7134	2.3774
	(.0347)	(−.0187)	(.0211)	(−.0577)
12, 18	−1.1641	1.3332	1.4957	2.4281
	(.1092)	(.0296)	(−.1966)	(−.0070)

Table 5. Richardson Extrapolation for the Binomial Distribution With p = .4

n_0, n_1	Percentiles 10	90	95	99
n = 50				
True values	−1.2776	1.2882	1.6694	2.3720
Normal approximation	−1.2816	1.2816	1.6449	2.3263
	(−.0040)	(−.0066)	(−.0245)	(−.0457)
	$n_0 + n_1 = 15$			
1, 14	−1.2692	1.2656	1.6423	2.4690
	(.0084)	(−.0226)	(−.0271)	(.0971)
3, 12	−1.1991	1.3197	1.6443	2.3799
	(.0785)	(.0315)	(−.0252)	(.0079)
4, 11	−1.2562	1.1647	1.6847	2.2989
	(.0214)	(−.1235)	(.0153)	(−.0731)
6, 9	−1.2007	1.0239	1.8705	2.4680
	(.0770)	(−.2643)	(.2010)	(.0961)
	$n_0 + n_1 = 20$			
2, 18	−1.2344	1.2776	1.6229	2.4184
	(.0432)	(−.0105)	(−.0466)	(.0464)
4, 16	−1.2823	1.2781	1.6526	2.3583
	(−.0047)	(−.0101)	(−.0168)	(−.0137)
5, 15	−1.2702	1.2816	1.6411	2.3911
	(.0074)	(−.0066)	(−.0283)	(.0191)
8, 12	−1.3054	1.2832	1.7722	2.5967
	(−.0278)	(−.0049)	(.1027)	(.2247)
n = 100				
True values	−1.2811	1.2892	1.6619	2.3475
Normal approximation	−1.2816	1.2816	1.6449	2.3263
	(−.0005)	(−.0076)	(−.0170)	(−.0212)
	$n_0 + n_1 = 20$			
2, 18	−1.2167	1.2576	1.5951	2.4224
	(.0644)	(−.0316)	(−.0668)	(.0749)
4, 16	−1.2738	1.2589	1.6383	2.3349
	(.0073)	(−.0303)	(−.0236)	(−.0126)
5, 15	−1.2674	1.2767	1.6183	2.4029
	(.0138)	(−.0125)	(−.0436)	(.0554)
8, 12	−1.3107	1.2668	1.7943	2.6437
	(−.0295)	(−.0224)	(.1324)	(.2962)
	$n_0 + n_1 = 30$			
3, 27	−1.2317	1.2922	1.6644	2.3581
	(.0494)	(.0030)	(.0025)	(.0106)
6, 24	−1.2379	1.2613	1.6580	2.3519
	(.0432)	(−.0279)	(−.0039)	(.0043)
7, 23	−1.2601	1.3154	1.6403	2.3348
	(.0210)	(.0262)	(−.0216)	(−.0128)
12, 18	−1.2439	1.2762	1.6676	2.3576
	(.0373)	(−.0130)	(.0057)	(.0101)

and

$$Q_n(u) = B_n(np + u\sqrt{np(1-p)}).$$

We approximated the percentiles $Q_n^{-1}(\alpha)$ for n, b, and λ as before, with $p = .2$ and .4. Note that the $\lambda = .75$ allocation seems to work best, but differs little from $\lambda = .8$ and .9. On the other hand, $\lambda = .6$ is poorer. (This is in agreement with our theory for class $\mathbf{D}_3$.) For $p = .2$, $n = 50$, 100, and $b = 15$, 20, the $\lambda = .75$ allocation does as well as the normal. For $b = 20$, 30 it is better, typically giving an additional significant figure. For $p = .4$, it is generally poorer, though far from terrible. This is understandable, since for $p = .5$, $A_1 = 0$, and the extrapolation is adding noise to the normal approximation.

In Table 6 we show the results for the bootstrap experiment. The population is $\sigma^2\chi_1^2$, and we are interested in a confidence bound for σ. We study the unadjusted bootstrap, that is, the percentiles of the bootstrap distribution of $(\bar{X}_n)^{1/2}$, where $\bar{X}$ is the sample mean. For $n = 50$, 100, and 500 we took 500 samples of size n from χ_1^2. For each sample we took 1,000 bootstrap samples and computed the .05, .1, and .95 percentiles of the bootstrap distribution of $(\bar{X}_n)^{1/2}$ for sample size n_0, n_1, and n. We study the behavior of the 90% lower confidence bound and the 90% confidence interval, that is, the .1 percentile and the interval between the .95 and .05 percentiles. This is Efron's (1979) percentile method, which we do not endorse in practice but use as a simple example of the bootstrap.

For each n we count the number of times the population parameter falls inside the confidence set, out of the 500 samples. We compute the average and standard deviation of the rescaled lower bound, that is, $\sqrt{n}(1 - G_n^{*-1}(.1))$, and the rescaled interval, that is, $I_n^*(.9) = \sqrt{n}(G_n^{*-1}(.95) - G_n^{*-1}(.05))$, where $G_n^{*-1}(\alpha)$ is the α percentile of the bootstrap distribution of $\bar{X}^{1/2}$. Table 6 shows clearly that Richardson extrapolation is a good approximation to the full bootstrap and is not very sensitive to the allocation of n_0 and n_1. The last entry gives estimated computation times on Sun workstations at the University of California. The expected linear saving in the sample size is confirmed.

APPENDIX: THEORY FOR EXAMPLE 2

We establish the claim asserted in Example 2 in the form of a theorem.

Theorem. $[an + b\sqrt{n}]$ is uniformly distributed (ud) mod 1 unless $b = 0$ and a is rational.

Proof. We refer repeatedly to the text of Kuipers and Niederreiter (KN 1974). Suppose that a is irrational. Note that

$$a(n+1) + b\sqrt{n+1} - an - b\sqrt{n} = a + bO(n^{-1/2}) \to a,$$

as $n \to \infty$. By theorem 3.3 of KN, $an + b\sqrt{n}$ is ud mod 1.

If a is rational we apply the following lemma.

Lemma. Let b_n be a sequence such that $\{b_{sj+k}\}_{j\geq 1}$ is ud mod 1 for $s \neq 0$ $(0 \leq k \leq s)$. Then if a is rational, $a = r/s$ and $an + b_n$ is ud mod 1.

Proof. Check Weyl's criterion (KN). Let $n = ms$. Then

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{l=1}^{n} \exp[2\pi i h(a_l + b_l)]\right| \\
&= \left|\frac{1}{ms}\sum_{j=0}^{m-1}\sum_{k=0}^{s-1} \exp\{2\pi i h[r(k/s) + b_{sj+k}]\}\right| \\
&\leq \frac{1}{s}\sum_{k=0}^{s-1}\left|\frac{1}{m}\sum_{j=0}^{m-1} \exp(2\pi i h b_{sj+k})\right| \to 0, \qquad \text{(A.1)}
\end{aligned}$$

as $m \to \infty$ by Weyl's criterion applied to $\{b_{sj+k}\}_{j\geq 1}$. If $n = ms + b$ $(0 < b < s)$, the difference from (A.1) is at most $b/ms \to 0$. The lemma follows by Weyl's criterion.

Let $b_n = b\sqrt{n}$. If $b > 0$, $b_{s(j+1)+k} - b_{sj+k}$ is decreasing to 0 in j, since $\sqrt{x}$ is concave. Moreover, $j(b_{s(j+1)+k} - b_{sj+k}) = \Omega(j^{1/2})$

Table 6. A Bootstrap Experiment

					Rescaled				
n	n_0	n_1	Lower-bound count	Interval count	Confidence-bound average	Confidence-bound SD	Average length	SD length	Time*
50	full	bootstrap	462	443	.83732	.007231	2.23093	.018770	1,603
50	2	18	455	439	.84251	.007652	2.26337	.019407	680
50	4	16	468	449	.83286	.007090	2.23915	.018084	680
100	full	bootstrap	457	445	.85825	.005957	2.25543	.015018	3,171
100	2	18	459	438	.85736	.006190	2.27469	.014191	688
100	4	16	472	446	.85685	.006639	2.26594	.014139	686
500	full	bootstrap	453	453	.89302	.003029	2.31675	.070418	15,754
500	5	45	454	448	.88568	.004092	2.31589	.009186	1,665
500	10	40	454	455	.89668	.004200	2.33916	.086705	1,666

NOTE: SD represents standard deviation.
* In central-processing-unit seconds.

$\to \infty$. By Fejer's theorem (KN, theorem 2.5), $\{b_{sj+k}\}$ is ud mod 1, and the theorem follows.

[*Received August 1986. Revised September 1987.*]

REFERENCES

Abramovitch, L., and Singh, K. (1985), "Edgeworth Corrected Pivotal Statistics and the Bootstrap," *The Annals of Statistics,* 13, 116–132.

Beran, R., and Millar, P. W. (1986), "Confidence Sets for a Multivariate Distribution," *The Annals of Statistics,* 14, 431–443.

Bhattacharya, R., and Ranga Rao, R. (1976), *Normal Approximation and Asymptotic Expansions,* New York: John Wiley.

Bickel, P. J., and Freedman, D. A. (1981), "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics,* 9, 1196–1217.

Efron, B. (1979), "Bootstrap Methods: Another Look at Jackknife," *The Annals of Statistics,* 7, 1–26.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications,* New York: John Wiley.

Isaacson, E., and Keller, H. B. (1966), *Analysis of Numerical Methods,* New York: John Wiley.

Kuipers, L., and Niederreiter, H. (1974), *Uniform Distribution or Sequences,* New York: John Wiley.

Singh, K. (1981), "On Asymptotic Accuracy of Efron's Bootstrap," *The Annals of Statistics,* 9, 1187–1195.

Statistica Sinica **7**(1997), 1-31

RESAMPLING FEWER THAN n OBSERVATIONS: GAINS, LOSSES, AND REMEDIES FOR LOSSES

P. J. Bickel, F. Götze and W. R. van Zwet

University of California, Berkeley,
University of Bielefeld and University of Leiden

Abstract: We discuss a number of resampling schemes in which $m = o(n)$ observations are resampled. We review nonparametric bootstrap failure and give results old and new on how the m out of n with replacement and without replacement bootstraps work. We extend work of Bickel and Yahav (1988) to show that m out of n bootstraps can be made second order correct, if the usual nonparametric bootstrap is correct and study how these extrapolation techniques work when the nonparametric bootstrap does not.

Key words and phrases: Asymptotic, bootstrap, nonparametric, parametric, testing.

1. Introduction

Over the last 10-15 years Efron's nonparametric bootstrap has become a general tool for setting confidence regions, prediction, estimating misclassification probabilities, and other standard exercises of inference when the methodology is complex. Its theoretical justification is based largely on asymptotic arguments for its consistency or optimality. A number of examples have been addressed over the years in which the bootstrap fails asymptotically. Practical anecdotal experience seems to support theory in the sense that the bootstrap generally gives reasonable answers but can bomb.

In a recent paper Politis and Romano (1994), following Wu (1990), and independently Götze (1993) showed that what we call the m out of n without replacement bootstrap with $m = o(n)$ typically works to first order both in the situations where the bootstrap works and where it does not.

The m out of n with replacement bootstrap with $m = o(n)$ has been known to work in all known realistic examples of bootstrap failure. In this paper,

- We show the large extent to which the Politis, Romano, Götze property is shared by the m out of n with replacement bootstrap and show that the latter has advantages.
- If the usual bootstrap works the m out of n bootstraps pay a price in efficiency. We show how, by the use of extrapolation the price can be avoided.

- We support some of our theory with simulations.

The structure of our paper is as follows. In Section 2 we review a series of examples of success and failure to first order (consistency) of (Efron's) nonparametric bootstrap (nonparametric). We try to isolate at least heuristically some causes of nonparametric bootstrap failure. Our framework here is somewhat novel. In Section 3 we formally introduce the m out of n with and without replacement bootstrap as well as what we call "sample splitting", and establish their first order properties restating the Politis-Romano-Götze result. We relate these approaches to smoothing methods. Section 4 establishes the deficiency of the m out of n bootstrap to higher order if the nonparametric bootstrap works to first order and Section 5 shows how to remedy this deficiency to second order by extrapolation. In Section 6 we study how the improvements of Section 5 behave when the nonparametric bootstrap doesn't work to first order. We present simulations in Section 7 and proofs of our new results in Section 8. The critical issue of choice of m and applications to testing will be addressed elsewhere.

2. Successes and Failure of the Bootstrap

We will limit our work to the i.i.d. case because the issues we discuss are clearest in this context. Extension to the stationary mixing case, as done for the m out of n without replacement bootstrap in Politis and Romano (1994), are possible but the study of higher order properties as in Sections 4 and 5 of our paper is more complicated.

We suppose throughout that we observe $X_1, \ldots, X_n$ taking values in $X = R^p$ (or more generally a separable metric space). i.i.d. according to $F \in \mathcal{F}_0$. We stress that $\mathcal{F}_0$ need not be and usually isn't the set of all possible distributions. In hypothesis testing applications, $\mathcal{F}_0$ is the hypothesized set, in looking at the distributions of extremes, $\mathcal{F}_0$ is the set of populations for which extremes have limiting distributions. We are interested in the distribution of a symmetric function of $X_1, \ldots, X_n$; $T_n(X_1, \ldots, X_n, F) \equiv T_n(\hat{F}_n, F)$ where $\hat{F}_n$ is defined to be the empirical distribution of the data. More specifically we wish to estimate a parameter which we denote $\theta_n(F)$, of the distribution of $T_n(\hat{F}_n, F)$, which we denote by $\mathcal{L}_n(F)$. We will usually think of θ_n as real valued, for instance, the variance of $\sqrt{n}$ median $(X_1, \ldots, X_n)$ or the 95% quantile of the distribution of $\sqrt{n}(\bar{X} - E_F(X_1))$.

Suppose $T_n(\cdot, F)$ and hence θ_n is defined naturally not just on $\mathcal{F}_0$ but on $\mathcal{F}$ which is large enough to contain all discrete distributions. It is then natural to estimate F by the nonparametric maximum likelihood estimate, (NPMLE), $\hat{F}_n$, and hence $\theta_n(F)$ by the plug in $\theta_n(\hat{F}_n)$. This is Efron's (ideal) nonparametric bootstrap. Since $\theta_n(F) \equiv \gamma(\mathcal{L}_n(F))$ and, in the cases we consider, computation of γ is straightforward the real issue is estimation of $\mathcal{L}_n(F)$. Efron's (ideal) bootstrap is to estimate $\mathcal{L}_n(F)$ by the distribution of $T_n(X_1^*, \dots, X_n^*, \hat{F}_n)$ where, given $X_1, \dots, X_n$ the X_i^* are i.i.d. $\hat{F}_n$, i.e. the bootstrap distribution of T_n. In practice, the bootstrap distribution is itself estimated by Monte Carlo or more sophisticated resampling schemes, (see DeCiccio and Romano (1989) and Hikley (1988)). We will not enter into this question further.

Theoretical analyses of the bootstrap and its properties necessarily rely on asymptotic theory, as $n \to \infty$ coupled with simulations. We restrict analysis to $T_n(\hat{F}_n, F)$ which are asymptotically stable and nondegenerate on $\mathcal{F}_0$. That is, for all $F \in \mathcal{F}_0$, at least weakly

$$\mathcal{L}_n(F) \to \mathcal{L}(F) \text{ non degenerate}$$
$$\theta_n(F) \to \theta(F) \tag{2.1}$$

as $n \to \infty$.

Using m out of n bootstraps or sample splitting implicitly changes our goal from estimating features of $\mathcal{L}_n(F)$ to features of $\mathcal{L}_m(F)$. This is obviously nonsensical without assuming that the laws converge.

Requiring non degeneracy of the limit law means that we have stabilized the scale of $T_n(\hat{F}_n, F)$. Any functional of $\mathcal{L}_n(F)$ is also a functional of the distribution of $\sigma_n T_n(\hat{F}_n, F)$ where $\sigma_n \to 0$ which also converges in law to point mass at 0. Yet this degenerate limit has no functional $\theta(F)$ of interest.

Finally, requiring that stability need occur only on $\mathcal{F}_0$ is also critical since failure to converge off $\mathcal{F}_0$ in a reasonable way is the first indicator of potential bootstrap failure.

2.1. When does the nonparametric bootstrap fail?

If θ_n does not depend on n, the bootstrap works, (is consistent on $\mathcal{F}_0$), if θ is continuous at all points of $\mathcal{F}_0$ with respect to weak convergence on $\mathcal{F}$. Conversely, the nonparametric bootstrap can fail if,

1. θ is not continuous on $\mathcal{F}_0$.
   An example we explore later is $\theta_n(F) = 1(F \text{ discrete})$ for which $\theta_n(\hat{F}_n)$ obviously fails if F is continuous.
   Dependence on n introduces new phenomena. In particular, here are two other reasons for failure we explore below.
2. θ_n is well defined on all of $\mathcal{F}$ but θ is defined on $\mathcal{F}_0$ only or exhibits wild discontinuities when viewed as a function on $\mathcal{F}$. This is the main point of examples 3-6.
3. $T_n(\hat{F}_n, F)$ is not expressible as or approximable on $\mathcal{F}_0$ by a continuous function of $\sqrt{n}(\hat{F}_n - F)$ viewed as an object weakly converging to a Gaussian limit in a suitable function space. (See Giné and Zinn (1989).) Example 7 illustrate this failure. Again this condition is a diagnostic and not necessary for failure as Example 6 shows.

We illustrate our framework and discuss prototypical examples of bootstrap success and failure.

2.2. Examples of bootstrap success

Example 1. Confidence intervals: Suppose $\sigma^2(F) \equiv \mathrm{Var}_F(X_1) < \infty$ for all $F \in \mathcal{F}_0$.
(a) Let $T_n(\hat{F}_n, F) \equiv \sqrt{n}(\bar{X} - E_F X_1)$. For the percentile bootstrap we are interested in $\theta_n(F) \equiv P_F[T_n(\hat{F}_n, F) \le t]$. Evidently $\theta(F) = \Phi(\frac{t}{\sigma(F)})$. In fact, we want to estimate the quantiles of the distribution of $T_n(\hat{F}_n, F)$. If $\theta_n(F)$ is the $1-\alpha$ quantile then $\theta(F) = \sigma(F) z_{1-\alpha}$ where z is the Gaussian quantile.
(b) Let $T_n(\hat{F}_n, F) = \sqrt{n}(\bar{X} - E_F X_1)/s$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. If $\theta_n(F) \equiv P_F(T_n(\hat{F}_n, F) \le t]$ then, $\theta(F) = \Phi(t)$, independent of F. It seems silly to be estimating a parameter whose value is known but, of course, interest now centers on $\theta'(F)$ the next higher order term in $\theta_n(F) = \Phi(t) + \frac{\theta'(F)}{\sqrt{n}} + O(n^{-1})$.

Example 2. Estimation of variance: Suppose F has unique median $m(F)$, continuous density $f(m(F)) > 0$, $E_F|X|^\delta < \infty$, some $\delta > 0$ for all $F \in \mathcal{F}_0$ and $\theta_n(F) = \mathrm{Var}_F(\sqrt{n}$ median $(X_1, \ldots, X_n))$. Then $\theta(F) = [4f^2(m(F))]^{-1}$ on $\mathcal{F}_0$.

Note that, whereas θ_n is defined for all empirical distributions F in both examples the limit $\theta(F)$ is 0 or ∞ for such distributions in the second. Nevertheless, it is well known (see Efron (1979)) that the nonparametric bootstrap is consistent in both examples in the sense that $\theta_n(\hat{F}_n) \xrightarrow{P} \theta(F)$ for $F \in \mathcal{F}_0$.

2.3. Examples of bootstrap failure

Example 3. Confidence bounds for an extremum: This is a variation on Bickel Freedman (1981). Suppose that all $F \in \mathcal{F}_0$ have a density f continuous and positive at $F^{-1}(0) > -\infty$. It is natural to base confidence bounds for $F^{-1}(0)$ on the bootstrap distribution of

$$T_n(\hat{F}_n, F) = n(\min_i X_i - F^{-1}(0)).$$

Let

$$\theta_n(F) = P_F[T_n(\hat{F}_n, F) > t] = (1 - F(\frac{t}{n} + F^{-1}(0))^n.$$

Evidently $\theta_n(F) \to \theta(F) = \exp(-f(F^{-1}(0))t)$ on $\mathcal{F}_0$.

The nonparametric bootstrap fails. Let

$$N_n^*(t) = \sum_{i=1}^n 1(X_i^* \le \frac{t}{n} + X_{(1)}), t > 0,$$

where $X_{(1)} \equiv \min_i X_i$ and $1(A)$ is the indicator of A. Given $X_{(1)}$, $n\hat{F}_n(\frac{t}{n} + X_{(1)})$ is distributed as 1+ binomial $(n-1, \frac{F(\frac{t}{n}+X_{(1)}) - F(X_{(1)})}{(1-F(X_{(1)}))})$ which converges weakly to a Poisson $(f(F^{-1}(0))t)$ variable. More generally, $n\hat{F}_n(\frac{\cdot}{n} + X_{(1)})$ converges weakly conditionally to $1 + N(\cdot)$, where N is a homogeneous Poisson process with parameter $f(F^{-1}(0))$. It follows that $N_n^*(\cdot)$ converges weakly (marginally) to a process $M(1 + N(\cdot))$ where M is a standard Poisson process independent of $N(\cdot)$. Thus if, in Efron's notation, we use P^* to denote conditional probability given $\hat{F}_n$ and let $\hat{F}_n^*$, be the empirical d.f. of $X_1^*, \ldots, X_n^*$ then $P^*[T_n(\hat{F}_n^*) > t] = P^*[N_n^*(t) = 0]$ converges weakly to the random variable $P[M(1+N(t)) = 0|N] = e^{-(N(t)+1)}$ rather than to the desired $\theta(F)$.

Example 4. Extrema for unbounded distributions: (Athreya and Fukuchi (1994), Deheuvels, Mason, Shorack (1993))

Suppose $F \in \mathcal{F}_0$ are in the domain of attraction of an extreme value distribution. That is: for some constants $A_n(F)$, $B_n(F)$,

$$n(1 - F)(A_n(F) + B_n(F)x) \to H(x, F),$$

where H is necessarily one of the classical three types (David (1981), p.259): $e^{-\beta x}1(\beta x \geq 0)$, $\alpha x^{-\beta}1(x \geq 0)$, $\alpha(-x)^{\beta}1(x \leq 0)$, for $\alpha, \beta \neq 0$. Let,

$$\theta_n(F) \equiv P[(\max(X_1, \ldots, X_n) - A_n(F))/B_n(F) \leq t] \to e^{-H(t,F)} \equiv \theta(F). \quad (2.2)$$

Particular choices of $A_n(F)$, for example, $F^{-1}(1 - \frac{1}{n})$ and $B_n(F)$ are of interest in inference. However, the bootstrap does not work. It is easy to see that

$$n(1 - \hat{F}_n(A_n(F) + tB_n(F))) \xrightarrow{w} N(t), \quad (2.3)$$

where N is an inhomogeneous Poisson process with parameter $H(t, F)$ and $\xrightarrow{w}$ denotes weak convergence. Hence if $T_n(\hat{F}_n, F) = (\max(X_1, \ldots, X_n) - A_n(F))/B_n(F)$ then

$$P^*[T_n(\hat{F}_n^*, F) \leq t] \overset{w}{\Rightarrow} e^{-N(t)}. \quad (2.4)$$

It follows that the nonparametric bootstrap is inconsistent for this choice of A_n, B_n. If it were consistent, then

$$P^*[T_n(\hat{F}_n^*, \hat{F}_n) \leq t] \xrightarrow{P} e^{-H(t,F)} \quad (2.5)$$

for all t and (2.5) would imply that it is possible to find random A real and $B \neq 0$ such that $N(Bt + A) = H(t, F)$ with probability 1. But $H(t, F)$ is continuous except at 1 point. So (2.4) and (2.5) contradict each other. Again, $\theta(F)$ is well defined for $F \in \mathcal{F}_0$ but not otherwise. Furthermore, small perturbations in F can lead to drastic changes in the nature of H, so that θ is not continuous if $\mathcal{F}_0$ is as large as possible.

Essentially the same bootstrap failure arises when we consider estimating the mean of distributions in the domain of attraction of stable laws of index $1 < \alpha \leq 2$. (See Athreya (1987))

Example 5. Testing and improperly centered U and V statistics: (Bretagnolle (1983))

Let $\mathcal{F}_0 = \{F : F[-c, c] = 1, E_F X_1 = 0\}$ and let $T_n(\hat{F}_n) = n\bar{X}^2 = n \int xy d\hat{F}_n(x) d\hat{F}_n(y)$. This is a natural test statistic for $H : F \in \mathcal{F}_0$. Can one use the nonparametric bootstrap to find the critical value for this test statistic? Intuitively, $\hat{F}_n \notin \mathcal{F}_0$ and this procedure is rightly suspect. Nevertheless, in more complicated contexts, it is a mistake made in practice. David Freedman pointed us to Freedman et al. (1994) where the Bureau of the Census appears to have fallen into such a trap. (see Hall and Wilson (1991) for other examples.) The nonparametric bootstrap may, in general, not be used for testing as will be shown in a forthcoming paper.

In this example, due to Bretagnolle (1983), we focus on $\mathcal{F}_0$ for which a general U or V statistic T is degenerate and show that the nonparametric bootstrap doesn't work. More generally, suppose $\psi : R^2 \to R$ is bounded and symmetric and let $\mathcal{F}_0 = \{F : \int \psi(x, y) dF(x) = 0 \text{ for all } y\}$.

Then, it is easy to see that

$$T_n(\hat{F}_n) = \int \psi(x, y) dW_n^0(x) dW_n^0(y), \tag{2.6}$$

where $W_n^0(x) \equiv \sqrt{n}(\hat{F}_n(x) - F(x))$ and well known that

$$\theta_n(F) \equiv P_F[T_n(\hat{F}_n) \le t] \to P\Big[\int \psi(xy) dW^0(F(x)) dW^0(F(y)) \le t\Big] \equiv \theta(F),$$

where W^0 is a Brownian Bridge. On the other hand it is clear that,

$$\begin{aligned} T_n(\hat{F}_n^*) &= n \int \psi(x, y) d\hat{F}_n^*(x) d\hat{F}_n(y) \\ &= \int \psi(x, y) dW_n^*(x) dW_n^{0*}(y) + 2 \int \psi(x, y) dW_n^0(x) dW_n^{0*}(y) \\ &\quad + \int \psi(x, y) dW_n^0(x) dW_n^0(y), \end{aligned} \tag{2.7}$$

where $W_n^{0*}(x) \equiv \sqrt{n}(\hat{F}_n^*(x) - \hat{F}_n(x))$. It readily follows that,

$$\begin{aligned} P^*[T_n(\hat{F}_n^*) \le t] \overset{w}{\Rightarrow} P\Big[&\int \psi(x, y) dW^0(F(x)) dW^0(F(y)) \\ &+2 \int \psi(x, y) dW^0(F(x)) d\tilde{W}^0(F(y)) \\ &+ \int \psi(x, y) d\tilde{W}^0(F(x)) d\tilde{W}^0(F(y)) \le t | \tilde{W}^0\Big], \end{aligned} \tag{2.8}$$

where $\tilde{W}^0, W^0$ are independent Brownian Bridges.

This is again an instance where $\theta(F)$ is well defined for $F \in \mathcal{F}$ but $\theta_n(F)$ does not converge for $F \notin \mathcal{F}_0$

Example 6. Nondifferentiable functions of the empirical: (Beran and Srivastava (1985) and Dümbgen (1993))

Let $\mathcal{F}_0 = \{F : E_F X_1^2 < \infty\}$ and

$$T_n(\hat{F}_n, F) = \sqrt{n}(h(\bar{X}) - h(\mu(F)))$$

when $\mu(F) = E_F X_1$. If h is differentiable the bootstrap distribution of T_n is, of course, consistent. But take $h(x) = |x|$, differentiable everywhere except at 0. It is easy to see then that if $\mu(F) \neq 0$, $\mathcal{L}_n(F) \to \mathcal{N}(0, \mathrm{Var}_F(X_1))$ but if $\mu(F) = 0$, $\mathcal{L}_n(F) \to |\mathcal{N}(0, \mathrm{Var}_F(X_1))|$.

The bootstrap is consistent if $\mu \neq 0$ but not if $\mu = 0$. We can argue as follows. Under $\mu = 0$, $\sqrt{n}(\bar{X}^* - \bar{X})$, $\sqrt{n}\bar{X}$ are asymptotically independent $\mathcal{N}(0, \sigma^2(F))$. Call these variables Z and Z'. Then, $\sqrt{n}(|\bar{X}^*| - |\bar{X}|) \overset{w}{\Rightarrow} |Z + Z'| - |Z'|$, a variable whose distribution is not the same as that of $|Z|$. The bootstrap distribution, as usual, converges (weakly) to the (random) conditional distribution of $|Z + Z'| - |Z'|$ given Z'. This phenomenon was first observed in a more realistic context by Beran and Srivastava (1985). Dümbgen (1993) constructs similar reasonable though more complicated examples where the bootstrap distribution never converges. If we represent $T_n(\hat{F}_n, F) = \sqrt{n}(T(\hat{F}_n) - T(F))$ in these cases then there is no linear $\dot{T}(F)$ such that $\sqrt{n}(T(\hat{F}_n) - T(F)) \approx \sqrt{n}\dot{T}(F)(\hat{F}_n - F)$ which permits the argument of Bickel-Freedman (1981).

2.4. Possible remedies

Putter and van Zwet (1993) show that if $\theta_n(F)$ is continuous for every n on $\mathcal{F}$ and there is a consistent estimate $\tilde{F}_n$ of F then bootstrapping from $\tilde{F}_n$ will work, i.e. $\theta_n(\tilde{F}_n)$ will be consistent except possibly for F in a "thin" set.

If we review our examples of bootstrap failure, we can see that constructing suitable $\tilde{F}_n \in \mathcal{F}_0$ and consistent is often a remedy that works for all $F \in \mathcal{F}_0$ not simply the complement of a set of the second category. Thus in Example 3 taking $\tilde{F}_n$ to be $\hat{F}_n$ kernel smoothed with bandwidth $h_n \to 0$ if $nh_n^2 \to 0$ works. In the first and simplest case of Example 4 it is easy to see, Freedman (1981), that taking $\tilde{F}_n$ as the empirical distribution of $X_i - \bar{X}$, $1 \le i \le n$ which has mean 0 and thus belongs to $\mathcal{F}_0$ will work. The appropriate choice of $\tilde{F}_n$ in the other examples of bootstrap failure is less clear. For instance, Example 4 calls for $\tilde{F}_n$ with estimated tails of the right order but how to achieve this is not immediate.

A general approach which we believe is worth investigating is to approximate $\mathcal{F}_0$ by a nested sequence of parametric models, (a sieve), $\{\mathcal{F}_{0,m}\}$, and use the M.L.E. $\tilde{\mathcal{F}}_{m(n)}$ for $\mathcal{F}_{0,m(n)}$, for a suitable sequence $m(n) \to \infty$. See Shen and Wong (1994) for example.

The alternative approach we study is to change θ_n itself as well as possibly its argument. The changes we consider are the m out of n with replacement bootstrap, the $(n-m)$ out of n jackknife or $\binom{n}{m}$ bootstrap discussed by Wu (1990) and Politis and Romano (1994), and what we call sample splitting.

3. The m Out of n Bootstraps

Let h be a bounded real valued function defined on the range of T_n, for instance, $t \to 1(t \leq t_0)$.

We view as our goal estimation of $\theta_n(F) \equiv E_F(h(T_n(\hat{F}_n, F)))$. More complicated functionals such as quantiles are governed by the same heuristics and results as those we detail below. Here are the procedures we discuss.

(i) *The n/n bootstrap* (*The nonparametric bootstrap*)

Let,

$$B_n(F) = E^* h(T_n(\hat{F}_n^*, F)) = n^{-n} \sum_{(i_1,\ldots,i_n)} h(T_n(X_{i_1}, \ldots, X_{i_n}, F)).$$

Then, $B_n \equiv B_n(\hat{F}_n) = \theta_n(\hat{F})$ is the n/n bootstrap.

(ii) *The m/n bootstrap*

Let

$$B_{m,n}(F) \equiv n^{-m} \sum_{(i_1,\ldots,i_m)} h(T_m(X_{i_1}, \ldots, X_{i_m}, F)).$$

Then, $B_{m,n} \equiv B_{m,n}(\hat{F}_n) = \theta_m(\hat{F}_n)$ is the m/n bootstrap.

(iii) *The $\binom{n}{m}$ bootstrap*

Let

$$J_{m,n}(F) = \binom{n}{m}^{-1} \sum_{i_1 < \cdots < i_m} h(T_m(X_{i_1}, \ldots, X_{i_m}, F)).$$

Then, $J_{m,n} \equiv J_{m,n}(\hat{F}_n)$ is the $\binom{n}{m}$ bootstrap.

(iv) *Sample splitting*

Suppose $n = mk$. Define,

$$N_{m,n}(F) \equiv k^{-1} \sum_{j=0}^{k-1} h(T_m(X_{jm+1}, \ldots, X_{(j+1)m}, F))$$

and $N_{m,n} \equiv N_{m,n}(\hat{F}_n)$ as the sample splitting estimates. For safety in practice one should start with a random permutation of the X_i.

The motivation behind $B_{m(n),n}$ for $m(n) \to \infty$ is clear. Since, by (2.1), $\theta_{m(n)}(F) \to \theta(F)$, $\theta_{m(n)}(\hat{F}_n)$ has as good a rationale as $\theta_n(\hat{F}_n)$. To justify $J_{m,n}$ note that we can write $\theta_m(F) = \theta_m(\underbrace{F \times \cdots \times F}_{m})$ since it is a parameter of the law of $T_m(X_1, \ldots, X_m, F)$. We now approximate $F \times \cdots \times F$ not by the m dimensional product measure $\underbrace{\hat{F}_n \times \cdots \times \hat{F}_n}_{m}$ but by sampling without replacement. Thus sample splitting is just k fold cross validation and represents a crude approximation to $\underbrace{F \times \cdots \times F}_{m}$.

The sample splitting method requires the least computation of any of the lot. Its obvious disadvantages are that it relies on an arbitrary partition of the sample and that since both m and k should be reasonably large, n has to be really substantial. This method and compromises between it and the $\binom{n}{m}$ bootstrap are studied in Blom (1976) for instance. The $\binom{n}{m}$ bootstrap differs from the m/n by $o_P(1)$ if $m = o(n^{1/2})$. Its advantage is that it never presents us with the ties which make resampling not look like sampling. As a consequence, as we note in Theorem 1, it is consistent under really minimal conditions. On the other hand it is somewhat harder to implement by simulation. We shall study both of these methods further, below, in terms of their accuracy.

A simple and remarkable result on $J_{m(n),n}$ has been obtained by Politis and Romano (1994), generalizing Wu (1990). This result was also independently noted and generalized by Götze (1993). Here is a version of the Götze result and its easy proof. Write J_m for $J_{m,n}, B_m$ for $B_{m,n}$, N_m for $N_{m,n}$.

Theorem 1. *Suppose* $\frac{m}{n} \to 0$, $m \to \infty$.
Then,

$$J_m(F) = \theta_m(F) + O_P((\frac{m}{n})^{\frac{1}{2}}). \tag{3.1}$$

If h is continuous and

$$T_m(X_1, \ldots, X_m, F) = T_m(X_1, \ldots, X_m, \hat{F}_n) + o_p(1) \tag{3.2}$$

then

$$J_m = \theta_m(F) + o_p(1). \tag{3.3}$$

Proof. Suppose T_m does not depend on F. Then, J_m is a U statistic with kernel $h(T_m(x_1, \ldots, x_m))$ and $E_F J_m = \theta_m(F)$ and (3.1) follows immediately. For (3.2) note that

$$\begin{aligned} & E_F|J_m - \binom{n}{m}^{-1} \sum_{i_1 < \cdots < i_m} h(T_m(X_{i_1}, \ldots, X_{i_m}, F))| \\ \leq \ & E_F|h(T_m(X_1, \ldots, X_m, \hat{F}_n)) - h(T_m(X_1, \ldots, X_m, F))| \end{aligned} \tag{3.4}$$

and (3.2) follows by bounded convergence. These results follows in the same way and even more easily for N_m. Note that if T_m does not depend on F, $E_F N_m = \theta_m(F)$ and,

$$\mathrm{Var}_F(N_m) = \frac{m}{n} \mathrm{Var}_F(h(T_m(X_1, \ldots, X_m))) > \mathrm{Var}_F(J_m). \tag{3.5}$$

Note. It may be shown, more generally under (3.2), that, for example, distances between the $\binom{n}{m}$ bootstrap distributions of $T_m(\hat{F}_m, F)$ and $\mathcal{L}_m(F)$ are also $O_P(m/n)^{1/2}$.

Let $X_j^{(i)} = (X_j, \dots, X_j)_{1\times i}$

$$h_{i_1,\dots,i_r}(X_1,\dots,X_r) = \frac{1}{r!}\sum_{1\le j_1\ne\cdots\ne j_r\le r} h(T_m(X_{j_1}^{(i_1)},\dots,X_{j_r}^{(i_r)},F)), \tag{3.6}$$

for vectors $\boldsymbol{i} = (i_1,\dots,i_r)$ in the index set

$$\Lambda_{r,m} = \{(i_1,\dots,i_r) : 1 \le i_1 \le \cdots \le i_r \le m, i_1+\cdots+i_r = m\}.$$

Then

$$B_{m,n}(F) = \sum_{r=1}^{m}\sum_{i\in\Lambda_{r,m}} \omega_{m,n}(\boldsymbol{i})\frac{1}{\binom{n}{r}}\sum_{1\le j_1\le\cdots\le j_r\le m} h_i(X_{j_1},\dots,X_{j_r},F), \tag{3.7}$$

where

$$\omega_{m,n}(\boldsymbol{i}) = \binom{n}{r}\binom{m}{i_1,\dots,i_r}/n^m.$$

Let

$$\theta_{m,n}(F) = E_F B_{m,n}(F) = \sum_{r=1}^{m}\sum_{i\in\Lambda_{r,m}} \omega_{m,n}(i)E_F h_i(X_1,\dots,X_r). \tag{3.8}$$

Finally, let

$$\delta_m(\frac{r}{m}) \equiv \max\{|E_F h_i(X_1,\dots,X_r) - \theta_m(F)| : \boldsymbol{i}\in\Lambda_{r,m}\} \tag{3.9}$$

and define $\delta_m(x)$ by extrapolation on $[0,1]$. Note that $\delta_m(1) = 0$.

Theorem 2. *Under the conditions of Theorem* 1

$$B_{m,n}(F) = \theta_{m,n}(F) + O_P(\frac{m}{n})^{\frac{1}{2}}. \tag{3.10}$$

If further,

$$\delta_m(1 - xm^{-1/2}) \to 0 \tag{3.11}$$

uniformly for $0 \le x \le M$*, all* $M < \infty$*, and* $m = o(n)$*, then*

$$\theta_{m,n}(F) = \theta_m(F) + o(1). \tag{3.12}$$

Finally if,

$$T_m(X_1^{(i_n)},\dots,X_r^{(i_r)},F) = T_m(X_1^{(i_1)},\dots,X_r^{(i_r)},\hat{F}_n) + o_P(1) \tag{3.13}$$

whenever $\boldsymbol{i} \in \Lambda_{r,m}, m \to \infty$ *and* $\max\{i_1, \ldots, i_r\} = O(m^{1/2})$ *then, if* $m \to \infty, m = o(n)$,

$$B_m = \theta_m(F) + o_p(1). \tag{3.14}$$

The proof of Theorem 2 will be given in the Appendix. There too we will show briefly that, in the examples we have discussed and some others, $J_{m(n)}$, $B_{m(n)}$, $N_{m(n)}$ are consistent for $m(n) \to \infty$, $\frac{m}{n} \to 0$.

According to Theorem 2, if T_n does not depend om F the m/n bootstrap works as well as the $\binom{n}{m}$ bootstrap if the value of T_m is not greatly affected by a number on the order of $\sqrt{m}$ ties in its argument. Some condition is needed. Consider $T_n(X_1, \ldots, X_n) = 1(X_i = X_j$ for some $i \neq j)$ and suppose F is continuous. The $\binom{n}{m}$ bootstrap gives $T_m = 0$ as it should. If $m \neq o(\sqrt{n})$ so that the $\binom{n}{m}$ and m/n bootstraps do not coincide asymptotically the m/n bootstrap gives $T_m = 1$ with positive probability. Finally, (3.13) is the natural extension of (3.2) and is as easy to verify in all our examples.

A number of other results are available for m out of n bootstraps.

Giné and Zinn (1989) have shown quite generally that when $\sqrt{n}(\hat{F}_n - F)$ is viewed as a member of a suitable Banach space $\mathcal{F}$ and,
(a) $T_n(X_1, \ldots, X_n, F) = t(\sqrt{n}(\hat{F}_n - F))$ for t continuous
(b) $\mathcal{F}$ is not too big
then B_n and $B_{m(n)}$ are consistent.

Praestgaard and Wellner (1993) extended these results to $J_{m(n)}$ with $m = o(n)$. Finally, under the Giné-Zinn conditions,

$$||\sqrt{m}(\hat{F}_n - F)|| = (\frac{m}{n})||\sqrt{n}(\hat{F}_n - F)|| = O_P(\frac{m}{n})^{1/2} \tag{3.15}$$

if $m = o(n)$. Therefore,

$$t(\sqrt{m}(\hat{F}_m - \hat{F}_n)) = t(\sqrt{m}(\hat{F}_m - F)) + o_p(1) \tag{3.16}$$

and consistency of N_m if $m = o(n)$ follows from the original Giné-Zinn result.

We close with a theorem on the parametric version of the m/n bootstrap which gives a stronger property than that of Theorem 1.

Let $\mathcal{F}_0 = \{F_\theta : \theta \in \Theta \subset R^p\}$ where Θ is open and the model is regular. That is, θ is identifiable, the F_θ have densities f_θ with respect to a σ finite μ and the map $\theta \to \sqrt{f_\theta}$ is continuously Hellinger differentiable with nonsingular derivative. By a result of LeCam (see Bickel, Klaassen, Ritov, Wellner (1993) for instance), there exists an estimate $\hat{\theta}_n$ such that, for all θ,

$$\int (f_{\hat{\theta}_n}^{1/2}(x) - f_\theta^{1/2}(x))^2 d\mu(x) = O_{P_\theta}(\frac{1}{n}). \tag{3.17}$$

Theorem 3. *Suppose $\mathcal{F}_0$ is as above. Let $F_\theta^m \equiv \underbrace{F_\theta \times \cdots \times F_\theta}_{m}$ and $\|\cdot\|$ denote the variational norm. Then*

$$\|F_{\hat{\theta}_n}^m - F_\theta^m\| = O_p((\frac{m}{n})^{1/2}). \tag{3.18}$$

Proof. This is consequence of the relations (LeCam (1986)).

$$\|F_{\theta_0}^m - F_{\theta_1}^m)\| \leq H(F_{\theta_0}^m, F_{\theta_1}^m)[(2 - H^2(F_{\theta_0}^m, F_{\theta_1}^m)], \tag{3.19}$$

where

$$H^2(F, G) = \frac{1}{2}\int(\sqrt{dF} - \sqrt{dG})^2 \tag{3.20}$$

and

$$H^2(F_{\theta_0}^m, F_{\theta_1}^m) = 1 - (\int \sqrt{f_{\theta_0} f_{\theta_1}} d\mu)^m = 1 - (1 - H^2(F_{\theta_0}, F))^m. \tag{3.21}$$

Substituting (3.21) into (3.20) and using (3.17) we obtain

$$\|F_{\hat{\theta}_n}^m - F_\theta^m\| = O_{P_\theta}(1 - \exp O_{P_\theta}(\frac{m}{n}))^{\frac{1}{2}}(1 + \exp O_{P_\theta}(\frac{m}{n})^{\frac{1}{2}}) = O_{P_\theta}(\frac{m}{n})^{\frac{1}{2}}. \tag{3.22}$$

This result is weaker than Theorem 1 since it refers only to the parametric bootstrap. It is stronger since even for $m = 1$, when sampling with and without replacement coincide, $\|\hat{F}_n - F_\theta\| = 1$ for all n if F_θ is continous.

4. Performance of B_m, J_m, and N_m as Estimates of $\theta_n(F)$

As we have noted, if we take $m(n) = o(n)$ then in all examples considered in which B_n is inconsistent, $J_{m(n)}$, $B_{m(n)}$, $N_{m(n)}$ are consistent. Two obvious questions are,

(1) How do we choose $m(n)$?

(2) Is there a price to be paid for using $J_{m(n)}$, $B_{m(n)}$, or $N_{m(n)}$ when B_n is consistent?

We shall turn to the first very difficult question in a forthcoming paper on diagnostics. The answer to the second is, in general, yes. To make this precise we take the point of view of Beran (1982) and assume that at least on $\mathcal{F}_0$,

$$\theta_n(F) = \theta(F) + \theta'(F)n^{-1/2} + O(n^{-1}), \tag{4.1}$$

where $\theta(F)$ and $\theta'(F)$ are regularly estimable on $\mathcal{F}_0$ in the sense of Bickel, Klaassen, Ritov and Wellner (1993) and $O(n^{-1})$ is uniform on Hellinger compacts. There are a number of general theorems which lead to such expansions. See, for example, Bentkus, Götze and van Zwet (1994).

Somewhat more generally than Beran, we exhibit conditions under which $B_n = \theta_n(\hat{F}_n)$ is fully efficient as an estimate of $\theta_n(F)$ and show that the m out n bootstrap with $\frac{m}{n} \to 0$ has typically relative efficiency 0.

We formally state a theorem which applies to fairly general parameters θ_n. Suppose ρ is a metric on $\mathcal{F}_0$ such that

$$\rho(\hat{F}_n, F_0) = O_{P_{F_0}}(n^{-1/2}) \text{ for all } F_0 \in \mathcal{F}_0. \tag{4.2}$$

Further suppose

A. $\theta(F), \theta'(F)$ are ρ Fréchet differentiable in $\mathcal{F}$ at $F_0 \in \mathcal{F}_0$. That is,

$$\theta(F) = \theta(F_0) + \int \psi(x, F_0)dF(x) + o(\rho(F, F_0)) \tag{4.3}$$

for $\psi \in L_2^0(F_0) \equiv \{h : \int h^2(x)dF_0(x) < \infty, \int h(x)dF_0(x) = 0\}$ and θ' obeys a similar identity with ψ replaced by another function $\psi' \in L_2^0(F_0)$. Suppose further

B. The tangent space of $\mathcal{F}_0$ at F_0 as defined in Bickel et al. (1993) is $L_2^0(F_0)$ so that ψ and ψ' are the efficient influence functions of θ, θ'. Essentially, we require that in estimating F there is no advantage in knowing $F \in \mathcal{F}_0$.

Finally, we assume,

C. For all $M < \infty$,

$$\sup\{|\theta_m(F) - \theta(F) - \theta'(F)m^{-1/2}| : \rho(F, F_0) \le M_n^{-1/2}, F \in \mathcal{F}\} = O(m^{-1}) \tag{4.4}$$

a strengthened form of (4.1). Then,

Theorem 4. *Under regularity of θ, θ' and A and C at F_0,*

$$\begin{aligned}\theta_m(\hat{F}_n) \;&\equiv\; \theta(F_0) + \theta'(F_0)m^{-1/2} + \frac{1}{n}\sum_{i=1}^{n}(\psi(X_i, F_0) + \psi'(X_i, F_0)m^{-1/2}) \\ &\quad + O(m^{-1}) + o_p(n^{-1/2}).\end{aligned} \tag{4.5}$$

If B also holds, $\theta_n(\hat{F}_n)$ is efficient. If in addition, $\theta'(F_0) \neq 0$, and $\frac{m}{n} \to 0$ the efficiency of $\theta_m(\hat{F}_n)$ is 0.

Proof. The expansions of $\theta(\hat{F}_n)\theta'(\hat{F}_n)$ are immediate by Fréchet differentiability and (4.5) follows by plugging these into (4.1). Since θ, θ' are assumed regular, ψ and ψ' are their efficient influence functions. Full efficiency of $\theta_n(\hat{F}_n)$ follows by general theory as given in Beran (1983) for special cases or by extending Theorem 2, p.63 of Bickel et al. (1993) in an obvious way. On the other hand, if $\theta'(F_0) \neq 0$, $\sqrt{n}(\theta_m(\hat{F}_n) - \theta_n(F_0))$ has asymptotic bias $(\sqrt{\frac{n}{m}} - 1)\theta'(F_0) + O(\frac{\sqrt{n}}{m}) = \sqrt{\frac{n}{m}}(1 + o(1))\theta'(F_0) \to \pm\infty$ and inefficiency follows.

Inefficiency results of the same type or worse may be proved about J_m and N_m but require going back to $T_m(X_1, \ldots, X_m, F)$ since J_m and B_n are not related in a simple way. We pursue this only by way of Example 1. If $\theta_n(F) = \mathrm{Var}_F(\sqrt{n}(\bar{X} - \mu(F)) = \theta(F)$, $B_m = B_n$ but,

$$J_m = \sigma^2(\hat{F}_n)(1 - \frac{m-1}{n-1}). \tag{4.6}$$

Thus, since $\theta'(F) = 0$ here, B_m is efficient but J_m has efficiency 0 if $\frac{m}{\sqrt{n}} \to \infty$. N_m evidently behaves in the same way.

It is true that the bootstrap is often used not for estimation but for setting confidence bounds. This is clearly the case for Example (1b), the bootstrap of t where $\theta(F)$ is known in advance. For example, Efron's percentile bootstrap uses the $(1-\alpha)$th quantile of the bootstrap distribution of $\bar{X}$ as a level $(1-\alpha)$ approximate upper confidence bound for μ. As is well known by now (see Hall (1992)), for example, this estimate although, when suitably normalized, efficiently estimating the $(1-\alpha)$th quantile of the distribution of $\sqrt{n}(\bar{X}-\mu)$ does not improve to order $n^{-1/2}$ over the coverage probability of the usual Gaussian based $\bar{X} + z_{1-\alpha}\frac{s}{\sqrt{n}}$. However, the confidence bounds based on the bootstrap distribution of the t statistic $\sqrt{n}(\bar{X}-\mu(F))/s$ get the coverage probability correct to order $n^{-1/2}$. Unfortunately, this advantage is lost if one were to use the $1-\alpha$ quantile of the bootstrap distribution of $T_m(\hat{F}_m, F) = \sqrt{m}(\bar{X}_m - \mu(F))/s_m$ where $\bar{X}_m$ and s_m^2 are the mean and usual estimate of variance bsed on a sample of size m. The reason is that, in this case, the bootstrap distribution function is

$$\Phi(t) - m^{-1/2}c(\hat{F}_n)\varphi(t)H_2(t) + O_P(m^{-1}) \tag{4.7}$$

rather than the needed,

$$\Phi(t) - n^{-1/2}c(\hat{F}_n)\varphi(t)H_2(t) + O_P(n{-}1).$$

The error committed is of order $m^{-1/2}$. More general formal results can be stated but we do not pursue this.

The situation for $J_{m(n)}$ and $N_{m(n)}$ which function under minimal conditions, is even worse as we discuss in the next section.

5. Remedying the Deficiencies of $B_{m(n)}$ when B_n is Correct: Extrapolation

In Bickel and Yahav (1988), motivated by considerations of computational economy, situations were considered in which θ_n has an expansion of the form (4.1) and it was proposed using B_m at $m = n_0$ and $m = n_1$, $n_0 < n_1 << n$ to produce estimates of θ_n which behave like B_n. We sketch the argument for a special case.

Suppose that, as can be shown for a wide range of situations, if $m \to \infty$,

$$B_m = \theta_m(\hat{F}_n) = \theta(\hat{F}_n) + \theta'(\hat{F}_n)m^{-1/2} + O_P(m^{-1}). \tag{5.1}$$

Then, if $n_1 > n_0 \to \infty$

$$\theta'(\hat{F}_n) = (B_{n_0} - B_{n_1})(n_0^{-1/2} - n_1^{-1/2})^{-1} + O_P(n_0^{-1/2}) \tag{5.2}$$

$$\theta(\hat{F}_n) = \frac{n_0^{-1/2}B_{n_1} - n_1^{-1/2}B_{n_0}}{n_0^{-1/2} - n_1^{-1/2}} + O_P(n_0^{-1}) \tag{5.3}$$

and hence a reasonable estimate of B_n is,

$$B_{n_0,n_1} \equiv \frac{n_0^{-1/2}B_{n_1} - n_1^{-1/2}B_{n_0}}{n_0^{-1/2} - n_1^{-1/2}} + \frac{(B_{n_0} - B_{n_1})}{n_0^{-1/2} - n_1^{-1/2}}n^{-1/2}.$$

More formally,

Proposition. *Suppose $\{\theta_m\}$ obey C of Section 4 and $n_0 n^{-1/2} \to \infty$. Then,*

$$B_{n_0,n_1} = B_n + o_p(n^{-1/2}). \tag{5.4}$$

Hence, under the conditions of Theorem 3 B_{n_0,n_1} is efficient for estimating $\theta_n(F)$.

Proof. Under C, (5.4) holds. By construction,

$$\begin{aligned} B_{n_0,n_1} &= \theta(\hat{F}_n) + \theta'(\hat{F}_n)n^{-1/2} + O_P(n_0^{-1}) + O_P(n_0^{-1/2}n^{-1/2}) \\ &= \theta_n(\hat{F}_n) + O_P(n_0^{-1}) + O_P(n_0^{-1/2}n^{-1/2}) + O_P(n^{-1}) \\ &= \theta_n(\hat{F}_n) + O_P(n_0^{-1}) \end{aligned} \tag{5.5}$$

and (5.4) follows.

Assorted variations can be played on this theme depending on what we know or assume about θ_n. If, as in the case where T_n is a t statistic, the leading term $\theta(F)$ in (4.1) is $\equiv \theta_0$ independent of F, estimation of $\theta(F)$ is unnecessary and we need only one value of $m = n_0$. We are led to a simple form of estimate, since ψ of Theorem 4 is 0,

$$\hat{\theta}_{n_0} = (1 - (\frac{n_0}{n})^{1/2})\theta_0 + (\frac{n_0}{n})^{1/2}B_{n_0}. \tag{5.6}$$

This kind of interpolation is used to improve theoretically the behaviour of B_{m_0} as an estimate of a parameter of a stable distribution by Hall and Jing (1993) though we argue below that the improvement is somewhat illusory.

If we apply (5.4) to construct a bootstrap confidence bound we expect the coverage probability to be correct to order $n^{-1/2}$ but the error is $O_P((n_0 n)^{-1/2})$ rather than $O_P(n^{-1})$ as with B_n. We do not pursue a formal statement.

5.1. Extrapolation of J_m and N_m

We discuss extrapolation for J_m and N_m only in the context of the simplest Example 1, where the essential difficulties become apparent and we omit general theorems.

In work in progress, Götze and coworkers are developing expansions for general symmetric statistics under sampling from a finite population. These results will permit general statements of the same qualitative nature as in our discussion of Example 1. Consider $\theta_m(F) = P_F[\sqrt{m}(\bar{X}_m - \mu(F)) \leq t]$. If $EX_1^4 < \infty$ and the X_i obey Cramér's condition, then

$$\theta_m(F) = \Phi(\frac{t}{\sigma(F)}) - K_3(F)\frac{\varphi}{6\sqrt{m}}(\frac{t}{\sigma(F)})H_2(\frac{t}{\sigma(F)}) + O(m^{-1}), \tag{5.7}$$

where $\sigma^2(F)$ and $K_3(F)$ are the second and third cumulants of F and $H_k(t) = \frac{(-1)^k}{\varphi(t)}\frac{d\varphi^k(t)}{dt^k}$. By Singh (1981), $B_m = \theta_m(\hat{F}_n)$ has the same expansion with F replaced by $\hat{F}_n$. However, by an easy extension of results of Robinson (1978) and Babu and Singh (1985),

$$J_m = \Phi(\frac{t}{\hat{K}_{2m}}) - \varphi(\frac{t}{\hat{K}_{2m}^{1/2}})\frac{\hat{K}_{3m}}{6m^{1/2}}H_2(\frac{t}{\hat{K}_{2m}^{1/2}}) + O_P(m^{-1}), \tag{5.8}$$

where

$$\hat{K}_{2m} = \sigma^2(\hat{F}_n)(1 - \frac{m-1}{n-1}) \tag{5.9}$$

$$\hat{K}_{3m} = K_3(\hat{F}_n)(1 - \frac{m-1}{n-1})(1 - \frac{2(m-1)}{n-2}). \tag{5.10}$$

The essential character of expansion (5.8), if $m/n = o(1)$, is

$$J_m = \theta(\hat{F}_n) + m^{-1/2}\theta'(\hat{F}_n) + \frac{m}{n}\gamma_n + O_P(m^{-1} + (\frac{m}{n})^2 + \frac{m^{\frac{1}{2}}}{n}), \tag{5.11}$$

where γ_n is $O_P(1)$ and independent of m. The m/n terms essentially come from the finite population correction to the variance and highter order cumulants of means of samples from a finite population. They reflect the obvious fact that if $m/n \to \lambda > 0$, J_m is, in general, incorrect even to first order. For instance, the variance of the $\binom{n}{m}$ bootstrap distribution corresponding to $\sqrt{m}(\bar{X} - \mu(F))$ is $1/n\sum(X_i - \bar{X})^2(1 - \frac{m-1}{n-1}))$ which converges to $\sigma^2(F)(1 - \lambda)$ if $m/n \to \lambda > 0$. What this means is that if expansions (4.1), (5.1) and (5.11) are valid, then using $J_{m(n)}$ again gives efficiency 0 compared to B_n. Worse is that (5.2) with J_{n_0}, J_{n_1} replacing B_{n_0}, B_{n_1} will not work since the n_1/n terms remain and make a contribution larger than $n^{-1/2}$ if $n_1/n^{1/2} \to \infty$. Essentially it is necessary to estimate the coefficient of m/n and remove the contribution of this term at the same time while keeping the three required values of m: $n_0 < n_1 < n_2$ such that the error $O(\frac{1}{n_0} + (\frac{n_2}{n})^2)$ is $o(n^{-1/2})$. This essentially means that n_0, n_1, n_2 have order larger than $n^{1/2}$ and smaller that $n^{3/4}$.

This effect persists if we seek to use an extrapolation of J_m for the t statistic. The coefficient of m/n as well as $m^{-1/2}$ needs to be estimated. An alternative here and perhaps more generally is to modify the t statistic being bootstrapped and extrapolated. Thus $T_m(X_1, \ldots, X_m, F) \equiv \sqrt{m}\frac{(\bar{X}_m - \mu(F))}{\hat{\sigma}(1-\frac{m-1}{n-1})^{1/2}}$ leads to an expansion for J_m of the form,

$$J_m = \Phi(t) + \theta'(\hat{F}_n) m^{-1/2} + O_P(m^{-1} + m/n), \tag{5.12}$$

and we again get correct coverage to order $n^{-1/2}$ by fitting the $m^{-1/2}$ term's coefficient, weighting it by $n^{-1/2} - m^{-1/2}$ and adding it to J_m.

If we know, as we sometimes at least suspect in symmetric cases, that $\theta(F) = 0$, we should appropriately extrapolate linearly in m^{-1} rather than $m^{-1/2}$.

The sample splitting situation is less satisfactory in the same example. Under (5.1), the coefficient of $1/\sqrt{m}$ is asymptotically constant. Put another way, the asymptotic correlation of B_m, $B_{\lambda m}$ as $m, n \to \infty$ for fixed $\lambda > 0$ is 1. This is also true for J_m under (5.11). However, consider N_m and N_{2m} (say) if $T_m = \sqrt{m}(\bar{X}_m - \mu(F))$. Let h be continuously boundedly differentiable, $n = 2km$. Then

$$\text{Cov}\,(N_m, N_{2m)} = \frac{1}{k}\,\text{Cov}\,\Big(h(m^{-1/2}(\sum_{j=1}^{m}(X_j - \bar{X}))), h((2m)^{-1/2}\sum_{j=1}^{2m}(X_j - \bar{X}))\Big). \tag{5.13}$$

Thus, by the central limit theorem,

$$\text{Corr}(N_m, N_{2m}) \to \frac{1}{2}\frac{\text{Cov}}{\text{Var}\,(Z_1)}\Big(h(Z_1), h\frac{(Z_1 + Z_2)}{\sqrt{2}}\Big), \tag{5.14}$$

where Z_1, Z_2 are independent Gaussian $\mathcal{N}(0, \sigma^2(F))$ and $\sigma^2(F) = \text{Var}_F(X_1)$. More generally, viewed as a process in m for fixed n, N_m centered and normalized is converging weakly to a non degenerate process. Thus, extrapolation does not make sense for N_m.

Two questions naturally present themselves.

(a) How do these games play out in practice rather than theory?

(b) If the expansions (5.1) and (5.11) are invalid beyond the 0th order, the usual situation when the nonparametric bootstrap is inconsistent, what price do we pay theoretically for extrapolation?

Simulations giving limited encouragement in response to question (a) are given in Bickel and Yahav (1988). We give some further evidence in Section 7. We now turn to question (b) in the next section.

6. Behaviour of the Smaller Resample Schemes When B_n is Inconsistent, and Presentation of Alternatives

The class of situations in which B_n does not work is too poorly defined for us to come to definitive conclusions. But consideration of the examples suggests the following,

A. When, as in Example 6, $\theta(F)$, $\theta'(F)$ are well defined and regularly estimable on $\mathcal{F}_0$ we should still be able to use extrapolation (suitably applied) to B_m and possibly to J_m to produce better estimates of $\theta_n(F)$.

B. When, as in all our other examples of inconsistency, $\theta(F)$ is not regularly estimable on $\mathcal{F}_0$ extrapolation should not improve over the behaviour of B_{n_0}, B_{n_1}.

C. If n_0, n_1 are comparable extrapolation should not do particularly worse either.

D. A closer analysis of T_n and the goals of the bootstrap may, in these "irregular" cases, be used to obtain procedures which should do better than the m/n or $\binom{n}{m}$ or extrapolation bootstraps.

The only one of these claims which can be made general is C.

Proposition 1. *Suppose*

$$B_{n_1} - \theta_n(F) \asymp B_{n_0} - \theta_n(F), \tag{6.1}$$

where $\asymp$ *indicates that the ratio tends to* 1*. Then, if* $n_0/n_1 \not\to 1$

$$B_{n_0,n_1} - \theta_n(F) \asymp B_{n_0} - \theta_n(F). \tag{6.2}$$

Proof. Evidently, $\frac{B_{n_0}+B_{n_1}}{2} = \theta_n(F) + \Omega(\epsilon_n)$ where $\Omega(\epsilon_n)$ means that the exact order of the remainder is ϵ_n. On the other hand,

$$\frac{B_{n_0} - B_{n_1}}{n_0^{-1/2} - n_1^{-1/2}}\Big(\frac{1}{\sqrt{n}} - \frac{1}{2}\big(\frac{1}{\sqrt{n_0}} + \frac{1}{\sqrt{n_1}}\big)\Big) = \Omega(\epsilon_n)\Big(\sqrt{\frac{n_0}{n}} + \Omega(1)\Big)$$

and the proposition follows.

We illustrate the other three claims in going through the examples.

Example 3. Here, $F^{-1}(0) = 0$,

$$\theta_n(F) = e^{f(0)t}\Big(1 + n^{-1}f'(0)\frac{t^2}{2}\Big) + O(n^{-2}) \tag{6.3}$$

which is of the form (5.1). But the functional $\theta(F)$ is not regular and only estimable at rate $n^{-1/3}$ if one puts a first order Lipschitz condition on $F \in \mathcal{F}_0$. On the other hand,

$$\begin{aligned}\log B_m &= m\log(1-\hat{F}_n(\frac{t}{m})) = m\log(1-(\hat{F}_n(\frac{t}{m})-\hat{F}_n(0)))\\ &= -m(F(\frac{t}{m})-F(0)) - \frac{m}{\sqrt{n}}\sqrt{n}(\hat{F}_n(\frac{t}{m})-F(\frac{t}{m})) + O_P(m(\hat{F}_n(\frac{t}{m})-F(\frac{t}{m}))^2)\\ &= tf(0)+\Omega(\frac{1}{m})+\Omega_P(\sqrt{\frac{m}{n}})+O_P(\frac{1}{n}), \qquad (6.4)\end{aligned}$$

where as before Ω, Ω_p indicate exact order. As Politis and Romano (1994) point out, $m = \Omega(n^{1/3})$ yields the optimal rate $n^{-1/3}$ (under f Lipschitz). Extrapolation does not help because the $\sqrt{\frac{m}{n}}$ term is not of the form $\gamma_n\sqrt{\frac{m}{n}}$ where γ_n is independent of m. On the contrary, as a process in m, $\sqrt{mn}(\hat{F}_n(\frac{t}{m})-F(\frac{t}{m}))$ behaves like the sample path of a stationary Gaussian process. So conclusion B holds in this case.

Example 4. A major difficulty here is defining $\mathcal{F}_0$ narrowly enough so that it is meaningful to talk about expansions of $\theta_n(F)$, $B_n(F)$ etc. If $\mathcal{F}_0$ in these examples is in the domain of attraction of stable laws or extreme value distributions it is easy to see that $\theta_n(F)$ can converge to $\theta(F)$ arbitrarily slowly. This is even true in Example 1 if we remove the Lipschitz condition on f. By putting on conditions as in Example 1, it is possible to obtain rates. Hall and Jing (1993) specify a possible family for the stable law attraction domain estimation of the mean mentioned in Example 4 in which $B_n = \Omega(n^{-\frac{1}{\alpha}})$ where α is the index of the stable law and α and the scales of the (assumed symmetric) stable distribution are not regularly estimable but for which rates such as $n^{-2/5}$ or a little better are possible. The expansions for $\theta_n(F)$ are not in powers of $n^{-1/2}$ and the expansion for B_n is even more complex. It seems evident that extrapolation does not help. Hall and Jing's (1993) theoretical results and simulations show that $B_{m(n)}$ though consistent, if $m(n)/n \to 0$, is a very poor estimate of $\theta_n(F)$. They obtain at least theoretically superior results by using interpolation between B_m and the, "known up to the value of the stable law index α", value of $\theta(F)$. However, the conditions defining $\mathcal{F}_0$ which permit them to deduce the order of B_n are uncheckable so that this improvement appears illusory.

Example 6. The discontinuity of $\theta(F)$ at $\mu(F) = 0$ under any reasonable specification of $\mathcal{F}_0$ makes it clear that extrapolation cannot succeed. The discontinuity in $\theta(F)$ persists even if we assume $\mathcal{F}_0 = \{\mathcal{N}(\mu, 1) : \mu \in R\}$ and use the parametric bootstrap. In the parametric case it is possible to obtain constant level confidence bounds by inverting the tests for $H : |\mu| = |\mu_0|$ vs $K : |\mu| > |\mu_0|$ using the noncentral χ_1^2 distribution of $(\sqrt{n}\bar{X})^2$. Asymptotically conservative confidence bounds can be constructed in the nonparametric case by forming a bootstrap confidence interval for $\mu(F)$ using $\bar{X}$ and then taking the image of this interval into $\mu \to |\mu|$. So this example illustrates points B and D.

We shall discuss claims A and D in the context of Example 5 or rather its simplest case with $T_n(\hat{F}_n, F) = n\bar{X}^2$. We begin with,

Proposition 2. *Suppose $E_F X_1^4 < \infty$, $E_F X_1 = 0$, and F satisfies Cramer's condition. Then,*

$$\begin{aligned} B_m \equiv\ & P^*[|\sqrt{m}\bar{X}^*|^2 \le t^2] = 2\Phi(\frac{t}{\hat{\sigma}}) - 1 - \frac{m\bar{X}^2}{\hat{\sigma}^3} t\varphi(\frac{t}{\hat{\sigma}}) - \frac{\hat{K}_3\bar{X}}{3\hat{\sigma}^4}\varphi H_3(\frac{t}{\hat{\sigma}}) \\ & + O_P(\frac{m}{n})^{3/2} + O_P(m^{-1}). \end{aligned} \tag{6.5}$$

If $m = \Omega(n^{1/2})$ then

$$P^*[|\sqrt{m}\bar{X}^*|^2 \le t^2] = P_F[n\bar{X}^2 \le t] + O_P(n^{-1/4}) \tag{6.6}$$

and no better choice of $\{m(n)\}$ is possible. If $n_0 < n_1$, $n_0 n^{-1/2} \to \infty$, $n_1 = o(n^{3/4})$,

$$B^{n_0,n_1} \equiv B_{n_0} - n_0\{(B_{n_1} - B_{n_0})/(n_1 - n_0)\} = P_F[n\bar{X}^2 \le t] + O_P(n^{-1/2}). \tag{6.7}$$

Proof. We make a standard application of Singh (1981). If $\hat{\sigma}^2 \equiv \frac{1}{n}\sum(X_i - \bar{X})^2$, $\hat{K}_3 \equiv \frac{1}{n}\sum(X_i - \bar{X})^3$ we get, after some algebra and Edgeworth expansion,

$$P^*[\sqrt{m}\bar{X}^* \le t] = \Phi\Big(\frac{t - \sqrt{m}\bar{X}}{\hat{\sigma}}\Big) - \frac{1}{\sqrt{m}}\varphi\Big(\frac{t-\sqrt{m}\bar{X}}{\hat{\sigma}}\Big)\frac{\hat{K}_3}{6}H_2\Big(\frac{t-\sqrt{m}\bar{X}}{\hat{\sigma}}\Big) + O_p(m^{-1}).$$

After Taylor expansion in $\sqrt{m}\frac{\bar{X}}{\hat{\sigma}}$ we conclude,

$$P^*[m\bar{X}_m^{*2} \le t^2] = 2\Phi(\frac{t}{\hat{\sigma}}) - 1 + \frac{\varphi'}{2}(\frac{t}{\hat{\sigma}})m\bar{X}^2 - \frac{\hat{K}_3}{3\hat{\sigma}^4}[\varphi H_3](\frac{t}{\hat{\sigma}})\bar{X} + O_P(\frac{m}{n})^{3/2} + O_P(m^{-1}) \tag{6.8}$$

and (6.5) follows. Since $m\bar{X}^2 = \Omega_P(m/n)$, (6.6) follows. Finally, from (6.5), if $n_0 n^{-1/2}$, $n_1 n^{-1/2} \to \infty$

$$\begin{aligned} B_{n_0} - n_0\{(B_{n_1} - B_{n_0})/(n_1 - n_0)\} &= 2\Phi(\frac{t}{\hat{\sigma}}) - 1 - \frac{K_3}{6}\varphi H_2(\frac{t}{\hat{\sigma}})\bar{X} + O_P(n^{-3/4}) \\ & + O_P(n^{-1/2}) + O_P(n^{-1/2}). \end{aligned} \tag{6.9}$$

Since $\bar{X} = O_P(n^{-1/2})$, (6.7) follows.

Example 5. As we noted, the case $T_n(\hat{F}_n, F) = n\bar{X}^2$ is the prototype of the use of the m/n bootstrap for testing discussed in Bickel and Ren (1995). From (6.7) of proposition 2 it is clear that extrapolation helps. However, it is not true that B^{n_0,n_1} is efficient since it has an unnecessary component of variance $(\hat{K}_3/6)[\varphi H_2](\frac{t}{\hat{\sigma}})\bar{X}$ which is negligible only if $K_3(F) = 0$. On the other hand it is easy to see that efficient estimation can be achieved by resampling not the X_i but the residuals $X_i - \bar{X}$, that is, a consistent estimate of F belonging to $\mathcal{F}_0$. So this example illustrates both A and D. Or in the general U or V statistic case, bootstrapping not $T_m(\hat{F}_n, F) \equiv n\int \psi(x,y)d\hat{F}_n(x)d\hat{F}_n(y)$ but rather $n\int \psi(x,y)d(\hat{F}_n - F)(x)d(\hat{F}_n - F)(y)$ is the right thing to do.

7. Simulations and Conclusions

The simulation algorithms were written and carried out by Adele Cutler and Jiming Jiang. Two situations were simulated, one already studied in Bickel and Yahav (1988) where the bootstrap is consistent (essentially Example 1) the other (essentially Example 3) where the bootstrap is inconsistent.

Sample size: $n = 50, 100, 400$

Bootstrap sample size: $B = 500$

Simulation size: $N = 2000$

Distributions: Example 1: $F = \chi_1^2$; Example 3: $F = \chi_2^3$

Statistics:

Example 1(a) modified: $T_m^{(a)} = \sqrt{m}(\sqrt{\bar{X}_m} - \sqrt{\mu(F)})$

Example 1(b): $T_m^{(b)} = \sqrt{m}\frac{(\bar{X}-\mu(F))}{s_m}$ where $s_m^2 = \frac{1}{m-1}\sum_{i=1}^m (X_i - \bar{X}_m)^2$.

Example 3. $T_m^{(c)} = m(\min(X_1, \ldots, X_m) - F^{-1}(0))$

Parameters of resampling distributions: $G_m^{-1}(.1)$, $G_m^{-1}(.9)$ where G_m is the distribution of T_m under the appropriate resampling scheme. We use B, J, N to distinguish the schemes m/n, $\binom{n}{m}$ and sample splitting respectively.

In Example 1 the G_m^{-1} parameters were used to form upper and lower "90%" confidence bounds for $\theta \equiv \sqrt{\mu(F)}$. Thus, from $T_m^{(a)}$,

$$\bar{\theta}_{mB} = \sqrt{\bar{X}_n} - \frac{1}{\sqrt{n}}G_{mB}^{-1}(.1)) \tag{7.1}$$

for the "90%" upper confidence bound based on the m/n bootstrap and, from $T_m^{(b)}$,

$$\bar{\theta}_{mB} = ((\bar{X}_n - \frac{s_n}{\sqrt{n}}G_{mB}^{-1}(.1))_+)^{1/2}, \tag{7.2}$$

where G_{mB} now corresponds to the t statistic. $\underline{\theta}_{mB}$, is defined similarly. The $\underline{\bar{\theta}}_{mJ}$ bounds are defined with G_{mJ} replacing G_{mB}. The $\underline{\bar{\theta}}_{mN}$ bounds are considered only for the unambiguous case m divides n and α an integer multiple of m/n.

Thus if $m = n/10$, $G^{-1}_{mN}(.1)$ is simply the smallest of the 10 possible values $\{T_m(X_{jm+1},\ldots,X_{(j+1)m},\hat{F}_n), 0 \le j \le 9\}$.

We also specify 2 subsample sizes $n_0 < n_1$ for the extrapolation bounds, $\underline{\theta}_{n_0,n_1}\bar{\theta}_{n_0,n_1}$. These are defined for $T^{(a)}_m$, for example, by.

$$\bar{\theta}_{n_0,n_1} = \sqrt{\bar{X}_n} - \frac{1}{\sqrt{n}}\Big\{\frac{(G^{-1}_{n_0B}(.1) + G^{-1}_{n_1B}(.1))}{2} + (n^{-1/2} - \frac{1}{2}(n_0^{-1/2} + n_1^{-1/2}))(G^{-1}_{n_0B}(.1) - G^{-1}_{n_1B}(.1))/(n_0^{-1/2} - n_1^{-1/2})\Big\}. \quad (7.3)$$

We consider roughly, $n_0 = 2\sqrt{n}, n_1 = 4\sqrt{n}$ and specifically, the triples (n, n_0, n_1): $(50, 15, 30), (100, 20, 40)$ and $(400, 40, 80)$.

In Example 3, we similarly study the lower confidence bound on $\theta = F^{-1}(0)$ given by,

$$\bar{\underline{\theta}}_m = \max(X_1,\ldots,X_n) - \frac{1}{n}G^{-1}_{mB}(.9). \quad (7.4)$$

and the extrapolation lower confidence bound

$$\underline{\theta}_{n_0,n_1} = \min(X_1,\ldots,X_n) - \frac{1}{n}\frac{(G^{-1}_{n_0B}(.9) + G^{-1}_{n_1B}(.9))}{2} + (n^{-1} - \frac{(n_0^{-1} + n_1^{-1})}{2})(G^{-1}_{n_0B}(.9) - G^{-1}_{n_1B}(.9))(n_0^{-1} - n_1^{-1}). \quad (7.5)$$

Note that we are using $1/m$ rather than $1/\sqrt{m}$ for extrapolation.

Measures of performance:

$CP \equiv$ Coverage probability, the actual probability under the situation simulated that the region prescribed by the confidence bound covers the true value of the parameter being estimated.

$$RMSE = \sqrt{E(\text{Bound}-\text{Actual quantile bound})^2}\ .$$

Here the actual quantile bound refers to what we would use if we knew the distribution of $T_n(X_1,\ldots,X_n,F)$. For example for $T^{(a)}_m$ we would replace $G^{-1}_{mB}(.1)$ in (7.1) for $F = \chi^2_1$ by the .1 quantile of the distribution of $\sqrt{n}(\sqrt{\frac{S_m}{m}} - 1)$ where S_m has a χ^2_m distribution, call it $G^{*-1}_m(.1)$. Thus, here,

$$MSE = \frac{1}{n}E(G^{-1}_{mB}(.1) - G^{*-1}_m(.1))^2.$$

We give in Table 1 results for the B_{n_1}, B_n and B_{n_0,n_1} bounds, based on $T^{(b)}_m$. The $T^{(a)}_m$ bootstrap, as in Bickel and Yahav (1988), has CP and $RMSE$ for

B_n, B_{n_0,n_1} and B_{n_1} agreeing to the accuracy of the Monte Carlo and we omit these tables.

We give the corresponding results for lower confidence bounds based on $T_m^{(c)}$ in Table 2. Table 3 presents results for sample splitting for $T_m^{(a)}$. Table 4 presents $T_m^{(a)}$ results for the $\binom{n}{m}$ bootstrap.

Table 1. The t bootstrap: Example 1(b) at 90% nominal level

n		Coverage probabilities (CP) B	B1	BR	$RMSE$ B	B1	BR
50							
	UB	.88	.90	.88	.19	.21	.19
	LB	.90	.90	.90	.15	.15	.15
100							
	UB	.90	.93	.89	.13	.14	.12
	LB	.91	.90	.91	.11	.10	.11
400							
	UB	.91	.94	.90	.06	.07	.06
	LB	.91	.90	.91	.05	.05	.05

Notes: (a) B1 corresponds to (6.2) or its LCB analogue for $m = n_1(n) = 30$, $40, 80$. Similarly B corresponds to $m = n$.
(b) BR corresponds to (6.3) or its LCB analogue with $(n_0, n_1) = (15, 30), (20, 40), (40, 80)$.

Table 2. The min statistic bootstrap: Example 3 at the nominal 90% level

n		CP	$RMSE$
50			
	B	.75	.01
	B1	.78	.07
	BR	.70	.07
	B1S	.82	.07
	BRS	.80	.07
100			
	B	.75	.04
	B1	.82	.03
	BR	.76	.04
	B1S	.87	.03
	BRS	.86	.03
400			
	B	.75	.09
	B1	.86	.01
	BR	.83	.01

Notes: (a) B corresponds to (6.4) with $m = n$, B1 with $m = n_1 = 30, 40, 80$, B1S with $m = n_1 = 16$.
(b) BR corresponds to (6.5) with $(n_0, n_1) = (15, 30), (20, 40), (40, 80)$, BRS with $(n_0, n_1) = (4, 16)$.

Table 3. Sample splitting in Example 1(a)

n		CP N	$B_{m(n)}$	$RMSE$ N	$B_{m(n)}$
50					
	UB	.82	.86	.32	.18
	LB	.86	.91	.28	.16
100					
	UB	.86	.89	.30	.14
	LB	.84	.90	.26	.12
400					
	UB	.85	.89	.28	.08
	LB	.86	.91	.27	.09

Note: N here refers to $m = .1n$ and $\alpha = .1$.

Table 4. The $\binom{n}{m}$ bootstrap and the m/n bootstrap in Example 1(a)

n	CP m	J	B	E(Length) J	B
50	16	.82	.88	.07	.09
100	16	.86	.88	.04	.05
400	40	.88	.90	.01	.01

Note: These figures are for simulation sizes of $N = 500$ and for 90% confidence intervals. Thus, the end points of the intervals are given by (7.1) and its UCB counterpart for B and J but with .1 replaced by .05. Similarly, $[E(\text{Bound}-\text{Actual quantile bound})^2]^{1/2}$ is replaced by the expected length of the confidence interval.

Conclusions. The conclusions we draw are limited by the range of our simulations. We opted for realistic sample sizes, of $50, 100$ and a less realistic 400. For $n = 50, 100$ the subsample sizes $n_1 = 30$ (for $n = 50$) and 40 (for $n = 100$) are of the order $n/2$ rather than $o(n)$. For all sample sizes $n_0 = 2\sqrt{n}$ is not really "of larger order than $\sqrt{n}$". The simulations in fact show the asymptotics as very good when the bootstrap works even for relatively small sample sizes. The story when the bootstrap doesn't work is less clear.

When the bootstrap works (Example 1)

- BR and B are very close both in terms of CP, and $RMSE$ even for $n = 50$ from Table 1.
- B1's CP though sometimes better than B's consistently differs more from B's and its $RMSE$ follows suit In particular, for UB in Table 1, the $RMSE$ of B1 is generally larger. LB exhibits less differences but this reflects that UB is

governed by the behaviour of χ_1^2 at 0. In simulations we do not present we get similar sharper differences for LB when F is a heavy tailed distribution such as Pareto with $EX^5 = \infty$

- The effects, however, are much smaller than we expected. This reflects that these are corrections to the coefficient of the $n^{-1/2}$ term in the expansion. Perhaps the most surprising aspect of these tables is how well B1 performs.
- From Table 3 we see that because the m we are forced to by the level considered is small, CP for the sample splitting bounds differs from the nominal level. If $n \to \infty$, $m/n \to .1$ the coverage probability doesn't tend to .1 since the estimated quantile doesn't tend to the actual quantile and both CP and $RMSE$ behave badly compared to B_m. This naive method can be fixed up (see Blom (1976) for instance). However, its simplicity is lost and the $\binom{n}{m}$ or m/n bootstrap seem preferable.
- The $\binom{n}{m}$ bounds are inferior as Table 4 shows. This reflects the presence of the finite population correction m/n, even though these bounds were considered for the more favorable sample size $m = 16$ for $n = 50, 100$ rather than $m = 30, 40$. Corrections such as those of Bertail (1994) or simply applying the finite population correction to s would probably bring performance up to that of B_{n_1}. But the added complication doesn't seem worthwhile.

When the bootstrap doesn't work (Example 3)

- From Table 2, as expected, the CP of the n/n bootstrap for the lower confidence bound was poor for all n. For $n_0 = 2\sqrt{n}, n_1 = 4\sqrt{n}$, CP for B1 was constantly better than B for all n. BR is worse than B1 but improves with n and was nearly as good as B1 for $n = 400$. For small n_0, n_1 both B1 and BR do much better. However, it is clear that the smaller m of B1S is better than all other choices.

We did not give results for the upper confidence bound because the granularity of the bootstrap distribution of $\min_i X_i$ for these values of m and n made $CP = 1$ in all cases.

Evidently, n_0, n_1 play a critical role here. What apparently is happening is that for n_0, n_1 not sufficiently small compared with n extrapolation picks up the wrong slope and moves the not so good B1 bound even further towards the poor B bound.

A message of these simulations to us is that extrapolation of the B_m plot may carry risks not fully revealed by the asymptotics. On the other hand, if n_0 and n_1 are chosen in a reasonable fashion extrapolation on the $\sqrt{n}$ scale works well when the bootstrap does. Two notes, based on simulations we do not present, should be added to the optimism of Bickel, Yahav (1988) however. There may be risk if n_0 is really small compared to $\sqrt{n}$. We obtained poor results for BR for the t statistics for $n_0 = 4$ and 2. Thus $n_0 = 4$, $n_1 = 16$ gave the wrong slope to the extrapolation which tended to overshoot badly. Also, taking n_1 and n_0 close to each other, as the theory of the 1988 paper suggests is appropriate for statistics possessing high order expansions when the expansion coefficients are deterministic, gives poor results. It can also be seen theoretically that the sampling variability of the bootstrap for m of the order $\sqrt{n}$ makes this prescription unreasonable.

The principal message we draw is that it is necessary to develop data driven methods of selection of m which lead to reasonable results over situations where both the bootstrap works and where it doesn't. Such methods are being pursued.

Acknowledgement

We are grateful to Jiming Jiang and Adele Cutler for essential programming, to John Rice for editorial comments, and to Kjell Doksum for the Blom reference. This research was supported by NATO Grant CRG 920650, Sonderforschungsbereich 343 Diskrete Strukturen der Mathematik, Bielefeld and NSA Grant MDA 904-94-H-2020.

Appendix

Proof of Theorem 2. For $\boldsymbol{i}=(i_1,\ldots,i_r) \in \Lambda_{r,m}$ let $U(\boldsymbol{i}) = \frac{1}{\binom{n}{r}} \sum\{h_i(X_{j_1},\ldots, X_{j_r},F) : 1 \le j_1 < \cdots < j_r \le n\}$. Then, since $h_{\boldsymbol{i}}$ as defined is symmetric in its arguments it is a U statistic and $\|h\|_\infty$ is an upper bound to its kernel. Hence

(a) $$\operatorname{Var}_F U(\boldsymbol{i}) \le \|h\|_\infty^2 \frac{r}{n}.$$ On the other hand,

(b) $$EU(\boldsymbol{i}) = E_F h_i(X_1,\ldots,X_r,F) \quad \text{and}$$

(c) $$B_{m,n}(F) = \sum_{r=1}^{m} \sum\{w_{m,n}(\boldsymbol{i}) U(\boldsymbol{i}) : \boldsymbol{i} \in \Lambda_{r,m}\}$$ by (3.7). Thus, by (c),

(d) $$\begin{aligned}\operatorname{Var}_F^{1/2} B_{m,n}(F) &\le \sum_{r=1}^{m} \sum\{w_{m,n}(\boldsymbol{i}) \operatorname{Var}_F^{1/2} U(\boldsymbol{i}) : \boldsymbol{i} \in \Lambda_{r,m}\} \\ &\le \max \operatorname{Var}_F^{1/2} U(\boldsymbol{i}) \le \|h\|_\infty \left(\frac{m}{n}\right)^{1/2}\end{aligned}$$

by (a). This completes the proof of (3.10).

The proof of (3.11) is more involved. By (3.8)

(e) $$|\theta_{m,n}(F) - \theta(F)| \le \sum_{r=1}^{m} \sum\{|E_F h_i(X_1,\ldots,X_r) - \theta_m(F)| w_{m,n}(\boldsymbol{i}) : \boldsymbol{i} \in \Lambda_{r,m}\}.$$

Let,

(f) $$P_{m,n}[R_m = r] = \sum\{w_{m,n}(\boldsymbol{i}) : \boldsymbol{i} \in \Lambda_{r,m}\}.$$

Expression (f) is easily recognized as the probability of getting $n-r$ empty cells when throwing n balls independently into m boxes without restrictions (see Feller (1968), p.19). Then it is well known or easily seen that

(g) $$E_{m,n}(R_m) = n(1-(1-\frac{1}{n})^m)$$

(h) $$\mathrm{Var}_{\,m,n}(R_m) = n\{(1-\frac{1}{n})^m - (1-\frac{2}{n})^m\} + n^2\{(1-\frac{2}{n})^m - (1-\frac{1}{n})^{2m}\}.$$

It is easy to check that, if $m = o(n)$

(i) $$E_{m,n}(R_m) = m(1+O(\frac{m}{n}))$$

(j) $$\mathrm{Var}_{\,m,n}(R_m) = O(m)$$

so that,

(k) $$\frac{R_m}{m} = 1 + O_P(m^{-1/2}).$$

From (e),

(l) $$|\theta_{m,n}(F) - \theta(F)| \le \sum_{r=1}^{m} \delta_m(\frac{r}{m}) P_{m,n}[R_m = r].$$

By (k), (l) and the dominated convergence theorem (3.12) follows from (3.11) and (k).

Finally, as in Theorem 1, we bound, as in (3.4),

(m) $$|B_{m,n}(F) - B_m(F)| \le \sum_{r=1}^{m} \sum\{E_F|h_{\boldsymbol{i}}(X_1,\dots,X_r) - h_{\boldsymbol{i}}(X_1,\dots,X_r,\hat{F}_n)| : \boldsymbol{i} \in \Lambda_{r,m}\} w_{m,n}(\boldsymbol{i}),$$

where

(n) $$h_{\boldsymbol{i}}(X_1,\dots,X_r,\hat{F}_n) = \frac{1}{r!} \sum_{1\le j_1 \ne \cdots \ne j_r \le r} h(T_m(X_{j_1}^{(i_1)},\dots,X_{j_r}^{(i_r)},\hat{F}_n)).$$

Let R_m be distributed according to (f) and given $R_m = r$, let $(I_1, \dots, I_r)$ be uniformly distributed on the set of partitions of m into r ordered integers, $I_1 \le I_2 \le \cdots \le I_r$. Then, from (m) we can write

$$|B_{m,n}(F) - B_m(F)| \le E\Delta(I_1, \dots, I_{R_m}), \tag{o}$$

where $\|\Delta\|_\infty \le \|h\|_\infty$. Further, by the continuity of h and (3.13), since $I_1 \le \cdots \le I_{R_m}$,

$$\Delta(I_1, \dots, I_{R_m})1(I_{R_m} \le \epsilon_m m) \xrightarrow{P} 0 \tag{p}$$

whenever $\epsilon_m = O(m^{-1/2})$. Now, $I_{R_m} > \epsilon_m m$,

$$m = \sum_{j=1}^{R_m} I_j \tag{q}$$

and $I_j \ge 1$ imply that,

$$m(1 - \epsilon_m) \ge \sum_{j=1}^{R_m - 1} I_j \ge (R_m - 1). \tag{r}$$

Thus,

$$P_{m,n}(I_{R_m} > \epsilon_m m) \le P_{m,n}(\frac{R_m}{m} - 1 \le -\epsilon_m + O(m^{-1})) \to 0 \tag{s}$$

if $\epsilon_m m^{1/2} \to \infty$. Combining (s), (k) and (p) we conclude that

$$E\Delta(I_1, \dots, I_{R_m}) \to 0 \tag{t}$$

and hence (o) implies (3.14).

The corollary follows from (e) and (f).

Note that this implies that the m/n bootstrap works if about $\sqrt{m}$ ties do not affect the value of T_m much.

Checking that J_m, B_m, N_m $m = o(n)$ works

The arguments we give for B_m also work for J_m only more easily since Theorem 1 can be verified. It is easier to directly verify that, in all our examples, the m/n bootstrap distribution of $T_n(\hat{F}_n, F)$ converges weakly (in probability) to its limit $\mathcal{L}(F)$ and conclude that Theorem 2 holds for all h continuous and bounded than to check the conditions of Theorem 2. Such verifications can be found in the papers we cite. We sketch in what follows how the conditions of Theorem 1 and 2 can be applied.

Example 1. (a) We sketch heuristically how one would argue for functionals considered in Section 2 rather than quantiles. For J_m we need only check that (2.6) holds since $\sqrt{m}(\bar{X} - \mu(F)) = o_p(1)$. For B_m note that the distribution of $m^{-1/2}(i_1 X_1 + \cdots + i_r X_r)$ differs from that of $m^{-1/2}(X_1 + \cdots + X_m)$ by $O(\sum_{j=1}^r \frac{(i_j^2-1)}{m})$. If we maximize $\sum_{j=1}^r (i_j^2 - 1)$ subject to $\sum_{j=1}^r i_j = m$, $i_j \geq 1$ we obtain $\frac{2(m-r)}{m} + \frac{(m-r)^2}{m}$. Thus for suitable $h, \delta_m(x) = 2(1-x) + \frac{1}{\sqrt{m}}(1-x)^2$ and the hypotheses of Theorem 2 hold.
(b) Note that,

$$P\Big[\sqrt{n}\frac{(\bar{X} - \mu(F))}{s} \leq t\Big] = P[\sqrt{n}(\bar{X} - \mu(F)) - st \leq 0]$$

and apply the previous arguments to $T_n(\hat{F}_n, F) \equiv \sqrt{n}(\bar{X} - \mu(F)) - st$.

Example 2. In Example 2 the variance corresponds to $h(x) = x^2$ if $T_m(\hat{F}_m, F) = m^{1/2}(\mathrm{med}(X_1, \ldots, X_m) - F^{-1}(\frac{1}{2}))$. An argument parallel to that in Efron (1979) works. Here is a direct argument for h bounded.

$$\text{(a)} \qquad P[\mathrm{med}(X_1^{(i_1)}, \ldots, X_r^{(i_r)}) \neq \mathrm{med}(X_1^{(i_1)}, \ldots, X_r^{(i_r-1)}, X_{r+1})] \leq \frac{1}{r+1}.$$

Thus,

$$\text{(b)} \qquad P[\mathrm{med}(X_1^{(i_1)}, \ldots, X_r^{(i_r)}) \neq \mathrm{med}(X_1, \ldots, X_m)] \leq \sum_{j=r+1}^{m} \frac{1}{j} \leq \log(\frac{m}{r}).$$

Hence for h bounded,

$$\delta_m(x) \leq \|h\|_\infty \log(\frac{1}{x})$$

and we can apply Theorem 2.

Example 3. Follows by checking (3.2) in Theorem 1 and that Theorem 2 applies for J_m by arguing as above for B_m. Alternatively, argue as in Athreya and Fukushi (1994).

Arguments similar to those given so far can be applied to the other examples.

References

Athreya, K. B. (1987). Bootstrap of the mean in the infinite variance case. *Ann. Statist.* **15**, 724-731.

Athreya, K. B. and Fukuchi, J. (1994). Bootstrapping extremes of I.I.D. random variables. Proceedings of Conference on Extreme Value Theory (NIST).

Babu, G. J. and Singh, K. (1985). Edgeworth expansions for sampling without replacement from finite populations. *J. Multivariate Anal.* **17**, 261-278.

Bentkus, V., Götze, F. and van Zwet, W. R. (1994). An Edgeworth expansion for symmetric statistics. Tech Report Univ. of Bielefeld.

Beran, R. (1982). Estimated sampling distributions: The bootstrap and competitors. *Ann. Statist.* **10**, 212-225.

Beran, R. and Srivastava, M. S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *Ann. Statist.* **13**, 95-115.

Bertail, P. (1994). Second order properties of an extrapolated bootstrap without replacement. Submitted to *Bernoulli.*

Bhattacharya, R. and Ghosh, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6**, 434-451.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9**, 1196-1217.

Bickel, P. J. and Ren, J. J. (1995). The m out of n bootstrap and goodness of fit tests with double censored data. Robust Statistics, Data Analysis and Computer Intensive Methods Ed. H. Rieder Lecture Notes in Statistics, Springer-Verlag.

Bickel, P. J., Klaassen, C. K., Ritov, Y. and Wellner, J. (1993). *Efficient and Adaptive Estimation in Semiparametric Models.* Johns Hopkins University Press, Baltimore.

Bickel, P. J. and Yahav, J. A. (1988). Richardson extrapolation and the bootstrap. J. *Amer. Statist. Assoc.* **83**, 387-393.

Blom, G. (1976). Some properties of incomplete U statistics. *Biometrika* **63**, 573-580.

Bretagnolle, J. (1981). Lois limites du bootstrap de certaines fonctionelles. *Ann. Inst. H. Poincaré, Ser.B* **19**, 281-296.

David, H. A. (1981). *Order Statistics.* 2nd edition, John Wiley, New York.

Deheuvels, P., Mason, D. and Shorack, G. (1993). Some results on the influence of extremes on the bootstrap. *Ann. Inst. H. Poincaré* **29**, 83-103.

DeCiccio T. J. and Romano, J. P. (1989). The automatic percentile method: Accurate confidence limits in parametric models, *Canad. J. Statist.* **17**, 155-169.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probab. Theory Related Fields* **95**, 125-140.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1-26.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, London, New York.

Feller W. (1968). *Probability Theory* v1. John Wiley, New York.

Freedman D. A. (1981). Bootstrapping regression models. *Ann. Statist.* **9**, 1218-1228.

Giné. E. and Zinn, J. (1989). Necessary conditions for the bootstrap of the mean. *Ann. Statist.* **17**, 684-691.

Götze, F. (1993). *Bulletin* I. M. S.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer Verlag, New York.

Hall, P. and Wilson, S. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**, 757-762.

Hall, P. and Jing B. Y. (1993). Performance of boostrap for heavy tailed distributions. Tech. Report A. N. U. Canberra.

Hinkley, D. V. (1988). Bootstrap methods (with discussion). *J. Roy. Statist. Soc. Ser.B* **50**, 321-337.

Mammen, E. (1992). When does bootstrap work? Springer Verlag, New York.

Politis, D. N. and Romano, J. P. (1994). A general theory for large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22**, 2031-2050.

Praestgaard, J. and Wellner, J. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21**, 2053-2086.

Putter, H. and van Zwet, W. R. (1993). Consistency of plug in estimators with applications to the bootstrap. Submitted to *Ann. Statist.*

Robinson, J. (1978). An asymptotic expansion for samples from a finite population. *Ann. Statist.* **6**, 1005-1011.

Shen, X. and Wong, W. (1994). Convergence rates of sieve estimates. *Ann. Statist.* **22**, 580-615.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9**, 1187-1195.

Wu, C. F. J. (1990). On the asymptotic properties of the jackknife histogram. *Ann. Statist.* **18**, 1438-1452.

Department of Statistics, University of California, Berkeley, 367 Evans Hall, #3860, Berkeley, CA 94720-3860, U. S. A.

Department of Mathematics, University of Bielefeld, Universitatsstrasse 4800, Bielefeld, Germany.

Department of Mathematics, University of Leiden, PO Box 9512 2300RA, Leiden, Netherlands.

(Received August 1995; accepted June 1996)

Chapter 7
High-Dimensional Statistics

Jianqing Fan

7.1 Contributions of Peter Bickel to Statistical Learning

7.1.1 Introduction

Peter J. Bickel has made far-reaching and wide-ranging contributions to many areas of statistics. This short article highlights his marvelous contributions to high-dimensional statistical inference and machine learning, which range from novel methodological developments, deep theoretical analysis, and their applications. The focus is on the review and comments of his six recent papers in four areas, but only three of them are reproduced here due to limit of the space.

Information and technology make data collection and dissemination much easier over the last decade. High dimensionality and large data sets characterize many contemporary statistical problems from genomics and neural science to finance and economics, which give statistics and machine learning opportunities with challenges. These relatively new areas of statistical science encompass the majority of the frontiers and Peter Bickel is certainly a strong leader in those areas.

In response to the challenge of the complexity of data, new methods and greedy algorithms started to flourish in the 1990s and their theoretical properties were not well understood. Among those are the boosting algorithms and estimation of insintric dimensionality. In 2005, Peter Bickel and his coauthors gave deep theoretical foundation on boosting algorithms (Bickel et al. 2005; Freund and Schapire 1997) and novel methods on the estimation of intrinsic dimensionality (Levina and Bickel 2005). Another example is the use of LASSO (Tibshirani 1996) for high-dimensional variable selection. Realizing issues with biases of the

J. Fan (✉)
Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, 08540, USA
e-mail: jqfan@princeton.edu

J. Fan et al. (eds.), *Selected Works of Peter J. Bickel*, Selected Works in Probability and Statistics 13, DOI 10.1007/978-1-4614-5544-8_7,

Lasso estimate, Fan and Li (2001) advocated a family of folded concave penalties, including SCAD, to ameliorate the problem and critically analyzed its theoretical properties including LASSO. See also Fan and Lv (2011) for further analysis. Candes and Tao (2007) introduced the Dantzig selector. Zou and Li (2008) related the family folded-concave penalty with the adaptive LASSO (Zou 2006). It is Bickel et al. (2009) who critically analyzed the risk properties of the Lasso and the Dantzig selector, which significantly helps the statistics and machine learning communities on better understanding various variable selection procedures.

Covariance matrix is prominently featured in many statistical problems from network and graphic models to statistical inferences and portfolio management. Yet, estimating large covariance matrices is intrinsically challenging. How to reduce the number of parameters in a large covariance matrix is a challenging issue. In Economics and Finance, motivated by the arbitrage pricing theory, Fan et al. (2008) proposed to use the factor model to estimate the covariance matrix and its inverse. Yet, the impact of dimensionality is still very large. Bickel and Levina (2008a,b) and Rothman et al. (2008) proposed the use of sparsity, either on the covariance matrix or precision matrix, to reduce the dimensionality. The penalized likelihood method used in the paper fits in the generic framework of Fan and Li (2001) and Fan and Lv (2011), and the theory developed therein is applicable. Yet, Rothman et al. (2008) were able to utilize the specific structure of the covariance matrix and Gaussian distribution to get much deeper results. Realizing intensive computation of the penalized maximum likelihood method, Bickel and Levina (2008a,b) proposed a simple threshold estimator that achieves the same theoretical properties.

The papers will be reviewed in chronological order. They have high impacts on the subsequent development of statistics, applied mathematics, computer science, information theory, and signal processing. Despite young ages of those papers, a google-scholar search reveals that these six papers have around 900 citations. The impacts to broader scientific communities are evidenced!

7.1.2 Intrinsic Dimensionality

A general consensus is that high-dimensional data admits lower dimensional structure. The complexity of the data structure is characterized by the intrinsic dimensionality of the data, which is critical for manifold learning such as local linear embedding, Isomap, Lapacian and Hessian Eigenmaps (Brand 2002; Donoho and Grimes 2003; Roweis and Saul 2000; Tenenbaum et al. 2000). These nonlinear dimensionality reduction methods go behond traditional methods such as principal component analysis (PCA), which deals only with linear projections, and multidimensional scaling, which focuses on pairwise distances.

The techniques to estimate the intrinsic dimensionality before Levina and Bickel (2005) are roughly two groups: eigenvalue methods or geometric methods. The former are based on the number of eigenvalues greater than a given threshold. They fail on nonlinear manifolds. While localization enhances the applicability of

PCA, local methods depend strongly on the choice of local regions and thresholds (Verveer and Duin 1995). The latter exploit the geometry of the data. A popular metric is the correlation dimension from fractal analysis. Yet, there are a couple of parameters to be tuned.

The main contributions of Levina and Bickel (2005) are twofolds: It derives the maximum likelihood estimate (MLE) from a statistical prospective and gives its statistical properties. The MLE here is really the local MLE in the terminology of Fan and Gijbels (1996). Before this seminal work, there are virtually no formal statistical properties on the estimation of intrinsic dimensionality. The methods were often too heuristical and framework was not statistical.

The idea in Levina and Bickel (2005) is creative and statistical. Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be a random sample in R^p. They are embedded in an m-dimensional space via $\mathbf{X}_i = g(\mathbf{Y}_i)$, with unknown dimensionality m and unknown functions g, in which $\mathbf{Y}_i$ has a smooth density f in R^m. Because of nonlinear embedding g, we can only use the local data to determine m. Let R be small, which asymptotically goes to zero. Given a point $\mathbf{x}$ in R^p, the local information is summarized by the number of observations falling in the ball $\{\mathbf{z} : \|\mathbf{z}-\mathbf{x}\| \le t\}$, which is denoted by $N_{\mathbf{x}}(t)$, for $0 \le t \le R$. In other words, the local information around $\mathbf{x}$ with radius R is characterized by the process

$$\{N_{\mathbf{x}}(t) : 0 \le t \le R\}. \tag{7.1}$$

Clearly, $N_{\mathbf{x}}(t)$ is a binomial distribution with number of trial n and probability of success

$$P(\|\mathbf{X}_i - \mathbf{x}\| \le t) \approx f(\mathbf{x})V(m)t^m, \quad \text{as } t \to 0, \tag{7.2}$$

where $V(m) = \pi^{m/2}[\Gamma(m/2+1)]^{-1}$ is the volume of the unit sphere in R^m. Recall that the approximation of the Binomial distribution by the Poison distribution. The process $\{N_{\mathbf{x}}(t) : 0 \le t \le R\}$ is approximately a Poisson process with the rate $\lambda(t)$, which is the derivative of (7.2), or more precisely

$$\lambda(t) = nf(\mathbf{x})V(m)mt^{m-1} \tag{7.3}$$

The parameters $\theta = \log f(\mathbf{x})$ and m can be estimated by the maximum likelihood using the local observation (7.1).

Assuming $\{N_{\mathbf{x}}(t), 0 \le t \le R\}$ is the inhomogeneous Poisson process with rate $\lambda(t)$. Then, the log-likelihood of observing the process is given by

$$L(m,\theta) = \int_0^R \log \lambda(t) dN_{\mathbf{x}}(t) - \int_0^R \lambda(t)dt. \tag{7.4}$$

This can be understood by breaking the data $\{N_{\mathbf{x}}(t), 0 \le t \le R\}$ as the data

$$\{N(\Delta), N(2\Delta)-N(\Delta), \cdots, N(T\Delta)-N(T\Delta-\Delta)\}, \qquad \Delta = R/T \tag{7.5}$$

with a large T and noticing that the data above follow independent poisson distributions with mean $\lambda(j\Delta)\Delta$ for the j-th increment (The dependence on $\mathbf{x}$

is suppressed for brevity of notation). Therefore, using the Poisson formula, the likelihood of data (7.5) is

$$\prod_{j=1}^{T} \exp(-\lambda(j\Delta)\Delta)[\lambda(j\Delta)\Delta]^{dN(j\Delta)}/(dN(j\Delta)!)$$

where $dN(j\Delta) = N(j\Delta) - N(j\Delta - \Delta)$. Taking the logarithm and ignoring terms independent of the parameters, the log-likelihood of the observing data in (7.5) is

$$\sum_{j=1}^{T} [\log \lambda(j\Delta)] dN(j\Delta) - \sum_{j=1}^{T} \lambda(j\Delta)\Delta.$$

Taking the limit as $\Delta \to 0$, we obtain (7.4).

By taking the derivatives with parameters m and θ in (7.4) and setting them to zero, it is easy to obtain that

$$\hat{m}_R(\mathbf{x}) = \left\{ \log(R) - N_{\mathbf{x}}(R)^{-1} \int_0^R (\log t) dN_{\mathbf{x}}(t) \right\}^{-1}. \tag{7.6}$$

Let $T_k(\mathbf{x})$ be the distance of the k-th nearest point to $\mathbf{x}$. Then,

$$\hat{m}_R(\mathbf{x}) = \left\{ N_{\mathbf{x}}(R)^{-1} \sum_{j=1}^{N_{\mathbf{x}}(R)} \log[R/T_j(\mathbf{x})] \right\}^{-1}. \tag{7.7}$$

Now, instead of fixing distance R, but fixing the number of points k, namely, taking $R = T_k(\mathbf{x})$ for a given k, then, $N_{\mathbf{x}}(R) = k$ by definition and the estimator becomes

$$\hat{m}_k(\mathbf{x}) = \left\{ k^{-1} \sum_{j=1}^{k} \log[T_k(\mathbf{x})/T_j(\mathbf{x})] \right\}^{-1}. \tag{7.8}$$

Levina and Bickel (2005) realized that the parameter m is global whereas the estimate $\hat{m}_k(\mathbf{x})$ is local, depending on the location $\mathbf{x}$. They averaged out the n estimates at the observed data points and obtained

$$\hat{m}_k = n^{-1} \sum_{i=1}^{n} \hat{m}_k(\mathbf{X}_i). \tag{7.9}$$

To reduce the sensitivity on the choice of the parameter k, they proposed to use

$$\hat{m} = (k_2 - k_1 + 1)^{-1} \sum_{k=k_1}^{k_2} \hat{m}_k \tag{7.10}$$

for the given choices of k_1 and k_2.

The above discussion reveals that the parameter m was estimated in a semi-parametric model in which $f(\mathbf{x})$ is fully nonparametric. Levina and Bickel (2005) estimates the global parameter m by averaging. Averaging reduces variances, but not biases. Therefore, it requires k to be small. However, when p is large, even with a small k, $T_k(\mathbf{x})$ can be large and so can be the bias. For semiparametric model, the work of Severini and Wong (1992) shows that the profile likelihood can have a better bias property. Inspired by that, an alternative version of the estimator is to use the global likelihood, which adds up the local likelihood (7.4) at each data point $\mathbf{X}_i$, i.e.

$$L(\theta_{\mathbf{x}_1},\cdots,\theta_{\mathbf{x}_n},m) = \sum_{i=1}^{n} L(\theta_{\mathbf{x}_i},m). \tag{7.11}$$

Following the same derivations as in Levina and Bickel (2005), we obtain the maximum profile likelihood estimator

$$\hat{m}_R^* = \left\{ [\sum_{i=1}^{n} N_{\mathbf{x}_i}(R)]^{-1} \sum_{i=1}^{n} \sum_{j=1}^{N_{\mathbf{x}_i}(R)} \log[R/T_j(\mathbf{x}_i)] \right\}^{-1}. \tag{7.12}$$

In its nearest neighbourhood form,

$$\hat{m}_k^* = \left\{ [n(k-2)]^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k} \log[T_k(\mathbf{x}_i)/T_j(\mathbf{x}_i)] \right\}^{-1}. \tag{7.13}$$

The reason for divisor $(k-2)$ instead of k is given in the next paragraph. It will be interesting to compare the performance of the method (7.13) with (7.9).

Levina and Bickel (2005) derived the asymptotic bias and variance of estimator (7.8). They advocated the normalization of (7.8) by $(k-2)$ rather than k. With this normalization, they derived that to the first order,

$$E(\hat{m}_k(\mathbf{x})) = m, \qquad \mathrm{var}(\hat{m}_k(\mathbf{x})) = m^2/(k-3). \tag{7.14}$$

The paper has huge impact on manifold learning with a wide range of applications from patten analysis and object classification to machine learning and statistics. It has been cited nearly 200 times within 6 years of publication.

7.1.3 Generalized Boosting

Boosting is an iterative algorithm that uses a sequence of weak classifiers, which perform slightly better than a random guess, to build a stronger learner (Freund 1990; Schapire 1990), which can achieve the Bayes error rate. One of successful boosting algorithms is the AdaBoost by Freund and Schapire (1997). The algorithm is powerful but appears heuristic at that time. It is Breiman (1998) who noted that the AdaBoost classifier can be viewed as a greedy algorithm for an empirical loss minimization. This makes a strong connection of the algorithm with statistical foundation that enables us to understand better theoretical properties.

Let $\{(\mathbf{X}_i, Y_i)\}_{i=1}^p$ be an i.i.d. sample where $Y_i \in \{-1, 1\}$. Let $\mathscr{H}$ be a set of weak learners. Breiman (1998) observed that the AdaBoost classifier is $\mathrm{sgn}(F(\mathbf{X}))$, where F is found by a greedy algorithm minimizing

$$n^{-1} \sum_{i=1}^{n} \exp(-Y_i F(\mathbf{X}_i)), \tag{7.15}$$

over the class of function

$$\mathscr{F}_\infty = \bigcup_{k=1}^{\infty} \{ \sum_{j=1}^{k} \lambda_j h_j : \lambda_j \in \mathbb{R}, h_j \in \mathscr{H} \}.$$

The work of Bickel et al. (2005) generalizes the AdaBoost in two important directions: more general class of convex loss functions and more flexible class of algorithms. This enables them to study the convergence of the algorithms and classifiers in a unified framework. Let us state in the population version of their algorithms to simplify the notation. The goal is to find $F \in \mathscr{F}_\infty$ to minimize $w(F) = EW(YF)$ for a convex loss $W(\cdot)$. They proposed two relaxed Guass-Southwell algorithms, which are basically coordinatewise optimization algorithms in high-dimensional space. Given the current value F_m and coordinate h, one intends to minimize $W(F_m + \lambda h)$ over $\lambda \in \mathbb{R}$. The first algorithm is as follows: For given $\alpha \in (0, 1]$ and F_0, find inductively $F_1, F_2, \ldots$, by $F_{m+1} = F_m + \lambda_m h_m$, $\lambda_m \in \mathbb{R}$, $h_m \in \mathscr{H}$ such that

$$W(F_{m+1}) \leq \alpha \min_{\lambda \in \mathbb{R}, h \in \mathscr{H}} W(F_m + \lambda h) + (1 - \alpha) W(F_m). \tag{7.16}$$

In particular, when λ_m and h_m minimize $W(F_m + \lambda h)$, then (7.16) is obviously satisfied with equality. The generalization covers the possibility that the minimum of $W(F_m + \lambda h)$ is not assumed or multiply assumed. The algorithm is very general in the sense that it does not even specify a way to find λ_m and h_m, but a necessary condition of (7.16) is that

$$W(F_{m+1}) \leq W(F_m).$$

In other words, the target value decreases each iteration. The second algorithm is the same as the first one but requires

$$W(F_{m+1}) + \gamma \lambda_m^2 \leq \alpha \min_{\lambda \in \mathbb{R}, h \in \mathscr{H}} [W(F_m + \lambda h) + \gamma \lambda^2] + (1 - \alpha) W(F_m). \tag{7.17}$$

Under such a broad class of algorithms, Bickel et al. (2005) demonstrated unambiguously and convincingly that the generalized boosting algorithm converges to the Bayes classifier. They further demonstrated that the generalized boosting algorithms are consistent when the sample versions are used. In addition, they were able to derive the algorithmic speed of convergence, minimax rates of the convergence of the generalized boosting estimator to the Bayes classifier, and the minimax rates of the Bayes classification regret. The results are deep and useful. The work puts boosting algorithms in formal statistical framework and provides insightful understanding on the fundamental properties of the boosting algorithms.

Regularization of Covariation Matrices

It is well known that the sample covariance matrix has unexpected features when p and n are of the same order (Johnstone 2001; Marčcenko and Pastur 1967). Regularization is needed in order to obtain the desire statistical properties. Peter Bickel pioneered the work on the estimation of large covariance and led the development of the field through three seminal papers in 2008. Before Bickel's work, the theoretical work is very limited, often confining the dimensionality to be finite [with exception of Fan et al. (2008)], which does not reflect the nature of high-dimensionality. It is Bickel's work that allows the dimensionality grows much faster than sample size.

To regularize the covariance matrices, one needs to impose some sparsity conditions. The methods to explore sparsity are thresholding and the penalized quasi-likelihood approach. The former is frequently applied to the situations in which the sparsity is imposed on the elements which are directly estimable. For example, when the $p \times p$ covariance matrix Σ is sparse, a natural estimator is the following thresholding estimator

$$\hat{\Sigma}_t = (\hat{\sigma}_{i,j} I(|\hat{\sigma}_{i,j}| \geq t)) \tag{7.18}$$

for a thresholding parameter t. Bickel and Levina (2008b) considered a class of matrix

$$\Big\{\Sigma : \sigma_{ii} \leq M, \quad \sum_{j=1}^{p} |\sigma_{ij}|^q \leq c_p, \forall i\Big\}, \tag{7.19}$$

for $0 \leq q < 1$. In particular, when $q = 0$, c_p is the maximum number of nonvanishing elements in each row. They showed that when the data follow the Gaussian distribution and $t_n = M'(n^{-1}(\log p))^{1/2}$ for a sufficiently large constant M',

$$\|\hat{\Sigma}_{t_n} - \Sigma\| = O_p\Big(c_p\left(n^{-1}\log p\right)^{(1-q)/2}\Big), \tag{7.20}$$

and

$$p^{-1}\|\hat{\Sigma}_{t_n} - \Sigma\|_F^2 = O_p\Big(c_p\left(n^{-1}\log p\right)^{1-q/2}\Big). \tag{7.21}$$

uniformly for the class of matrices in (4.3), where $\|\mathbf{A}\|^2 = \lambda_{\max}(\mathbf{A}^T\mathbf{A})$ is the operator norm of a matrix $\mathbf{A}$ and $\|\mathbf{A}\|_F^2 = \sum_{i,j} a_{ij}^2$ is the Frobenius norm. Similar results were derived when the distributions are sub-Gaussian or have finite moments or when t_n is chosen by cross-validation which is very technically challenging and novel. This along with Bickel and Levina (2008b) and El Karoui (2008) are the first results of this kind, allowing $p \gg n$, as long as c_p does not grow too fast.

When the covariance matrix admits a banded structure whose off-diagonal elements decay quickly:

$$\sum_{j:|i-j|>k} |\sigma_{ij}| \leq Ck^{-\alpha}, \qquad \forall i \text{ and } k, \tag{7.22}$$

as arising frequently in time-series application including the covariance matrix of a weak-dependent stationary time series, Bickel and Levina (2008a) proposed a banding or more generally tapering to take advantage of prior sparsity structure. Let

$$\hat{\Sigma}_{B,k} = (\hat{\sigma}_{ij} I(|i-j| \leq k)$$

be the banded sample covariance matrix. They showed that by taking $k_n \asymp (n^{-1}\log p)^{-1/(2(\alpha+1))}$,

$$\|\hat{\Sigma}_{B,k_n} - \hat{\Sigma}\| = O_p\left[(n^{-1}\log p)^{\alpha/(2(\alpha+1))}\right] = \|\hat{\Sigma}_{B,k_n}^{-1} - \hat{\Sigma}^{-1}\| \tag{7.23}$$

uniformly in the class of matrices (7.22) with additional restrictions that

$$c \leq \lambda_{\min}(\Sigma) \leq \lambda_{max}(\Sigma) \leq C.$$

This again shows that large sparse covariance matrix can well be estimated even when $p \geq n$. The results are related to the estimation of spectral density (Fan and Gijbels 1996), but also allow non-stationary covariance matrices.

When the precision matrix $\Omega = \Sigma^{-1}$ is sparse, there is no easy way to apply thresholding rule. Hence, Rothman et al. (2008) appealed to the penalized likelihood method. Let $\ell_n(\theta)$ be the quasi-likelihood function based on a sample of size n and it is known that θ is sparse. Then, the penalized likelihood admits the form

$$\ell_n(\theta) + \sum_j p_\lambda(|\theta_j|). \tag{7.24}$$

Fan and Li (2001) advocated the use of folded-concave penalty p_λ to have a better bias property and put down a general theory. In particular, when the data $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are i.i.d. from $N(0, \Sigma)$, the penalized likelihood reduces to

$$\mathrm{tr}(\Omega\hat{\Sigma}) - \log|\Omega| + \sum_{i,j} p_\lambda(|\omega_{ij}|), \tag{7.25}$$

where the matrix Ω is assumed to be sparse and is of primary interest. Rothman et al. (2008) utilized the fact that the diagonal elements are non-vanishing and should not be penalized. They proposed the penalized likelihood estimator $\hat{\Omega}_\lambda$, which maximizes

$$\mathrm{tr}(\Omega\hat{\Sigma}) - \log|\Omega| + \lambda \sum_{i \neq j} |\omega_{ij}|. \tag{7.26}$$

They showed that when $\lambda \asymp [(\log p)/n]^{1/2}$,

$$\|\hat{\Omega}_\lambda - \Omega\|_F^2 = O_P\left(\sqrt{\frac{(p+s)(\log p)}{n}}\right), \tag{7.27}$$

where s is the number of nonvanishing off diagonal elements. Note that there are $p+2s$ nonvanishing elements in Ω and (7.27) reveals that each nonsparse element is estimated, on average, with rate $(n^{-1}(\log p))^{-1/2}$.

Note that thresholding and banding are very simple and easy to use. However, they are usually not semi-definite. Penalized likelihood can be used to enforce the positive definiteness in the optimization. It can also be applied to estimate sparse covariance matrices and sparse Chelosky decomposition; see Lam and Fan (2009).

The above three papers give us a comprehensive overview on the estimability of large covariance matrices. They have inspired many follow up work, including Levina et al. (2008), Lam and Fan (2009), Rothman et al. (2009), Cai et al. (2010), Cai and Liu (2011), and Cai and Zhou (2012), among others. In particular, the work inspires Fan et al. (2011) to propose an approximate factor model, allowing the idiosyncratic errors among financial assets to have a sparse covariance matrix, that widens significantly the scope and applicability of the strict factor model in finance. It also helps solving the aforementioned semi-definiteness issue, due to thresholding.

7.1.4 Variable Selections

Peter Bickel contributions to high-dimensional regression are highlighted by his paper with Ritov and Tsybakov (Bickel et al. 2009) on the analysis of the risk properties of the LASSO and Dantzig selector. This is done in least-squares setting on the nonparametric regression via basis approximations (approximate linear model) or linear model itself. This is based the following important observations in Bickel et al. (2009).

Recall that the LASSO estimator $\hat{\beta}_L$ minimizes

$$(2n)^{-1}\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{7.28}$$

A necessary condition is that 0 belongs to the subgradient of the function (7.28), which is the same as

$$\|n^{-1}\mathbf{X}(\mathbf{Y}-\mathbf{X}\hat{\beta}_L)\|_\infty \le \lambda. \tag{7.29}$$

The Danzig selector (Candes and Tao 2007) is defined by

$$\hat{\beta}_D = \operatorname{argmin}\left\{\|\beta\|_1 : \|n^{-1}\mathbf{X}(\mathbf{Y}-\mathbf{X}\beta)\|_\infty \le \lambda\right\}. \tag{7.30}$$

Thus, $\hat{\beta}_D$ satisfies (7.29), having a smaller L_1-norm than LASSO, by definition. They also show that for both the Lasso and the Danzig estimator, their estimation error δ satisfies

$$\|\delta_{J^c}\|_1 \le c\|\delta_J\|_1$$

with probability close to 1, where J is the subset of non-vanishing true regression coefficients. This leads them to define restricted eigenvalue assumptions.

For linear model, Bickel et al. (2009) established the convergence rates of

$$\|\hat{\beta}_D - \beta\|_p \text{ for } p \in [1,2] \quad \text{and} \quad \|\mathbf{X}(\hat{\beta}_D - \beta)\|_2. \tag{7.31}$$

The former is on the convergence rate of the estimator and the latter is on the prediction risk of the estimator. They also established the rate of convergence for the Lasso estimator. Both estimators admit the same rate of convergence under the same conditions. Similar results hold when the method is applied to nonparametric regression. This leads Bickel et al. (2009) to conclude that both the Danzig selector and Lasso estimator are equivalent.

The contributions of the paper are multi-fold. First of all, it provides a good understanding on the performance of the newly invented Danzig estimator and its relation to the Lasso estimator. Secondly, it introduced new technical tools for the analysis of penalized least-squares estimator. Thirdly, it derives various new results, including oracle inequalities, for the Lasso and the Danzig selector in both linear model and nonparametric regression model. The work has a strong impact on the recent development of the high-dimensional statistical learning. Within 3 years of its publications, it has been cited around 300 times!

References

Bickel PJ, Levina E (2008a) Regularized estimation of large covariance matrices. Ann Stat 36: 199–227

Bickel PJ, Levina E (2008b) Covariance regularization by thresholding. Ann Stat 36:2577–2604

Bickel PJ, Ritov Y, Zakai A (2005) Some theory for generalized boosting algorithms. J Mach Learn Res 7:705–732

Bickel PJ, Ritov Y, Tsybakov A (2009) Simultaneous analysis of Lasso and Dantzig selector. Ann Statist 37:1705–1732.

Brand M (2002) Charting a manifold. In: Advances in NIPS, vol 14. MIT Press, Cambridge, MA
Breiman L (1998) Arcing classifiers (with discussion). Ann Stat 26:801–849
Cai T, Liu W (2011) Adaptive thresholding for sparse covariance matrix estimation. J Am Stat Assoc 494:672–684
Cai T, Zhou H (2012) Minimax estimation of large covariance matrices under ℓ_1 norm (with discussion). Stat Sin, to appear
Cai T, Zhang C-H, Zhou H (2010) Optimal rates of convergence for covariance matrix estimation Ann Stat 38:2118–2144
Candes E, Tao T (2007) The Dantzig selector: statistical estimation when p is much larger than n (with discussion). Ann Stat 35:2313–2404
Donoho DL, Grimes C (2003) Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. Technical Report TR 2003-08, Department of Statistics, Stanford University, 2003
El Karoui N (2008) Operator norm consistent estimation of large-dimensional sparse covariance matrices. Ann Stat 36:2717–2756
Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman and Hall, London
Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360
Fan J, Lv J (2011) Non-concave penalized likelihood with NP-dimensionality. IEEE Inf Theory 57:5467–5484
Fan J, Fan Y, Lv J, (2008) Large dimensional covariance matrix estimation via a factor model. J Econ 147:186–197
Fan J, Liao Y, Mincheva M (2011) High dimensional covariance matrix estimation in approximate factor models. Ann Statist 39:3320–3356
Freund Y (1990) Boosting a weak learning algorithm by majority. In: Proceedings of the third annual workshop on computational learning theory. Morgan Kaufmann, San Mateo
Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55:119–139
Johnstone IM (2001) On the distribution of the largest eigenvalue in principal components analysis. Ann Stat 29:295–327
Lam C, Fan J (2009) Sparsistency and rates of convergence in large covariance matrices estimation. Ann Stat 37:4254–4278
Levina E, Bickel PJ (2005) Maximum likelihood estimation of intrinsic dimension. In: Saul LK, Weiss Y, Bottou L (eds) Advances in NIPS, vol 17. MIT Press, Cambridge, MA
Levina E, Rothman AJ, Zhu J (2008) Sparse estimation of large covariance matrices via a nested lasso penalty. Ann Stat Appl Stat 2:245–263
Marčcenko VA, Pastur LA (1967) Distributions of eigenvalues of some sets of random matrices. Math USSR-Sb 1:507–536
Rothman AJ, Bickel PJ, Levina E, Zhu J (2008) Sparse permutation invariant covariance estimation. Electron J Stat 2:494–515
Rothman AJ, Levina E, Zhu J (2009) Generalized thresholding of large covariance matrices. J Am Stat Assoc 104:177–186
Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326
Schapire R (1990) Strength of weak learnability. Mach Learn 5:197–227
Severini TA, Wong WH (1992) Generalized profile likelihood and conditional parametric models. Ann Stat 20:1768–1802
Tenenbaum JB, de Silva V, Landford JC(2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323

Tibshirani R (1996) Regression shrinkage and selection via lasso. J R Stat Soc B 58:267–288
Verveer P, Duin R (1995) An evaluation of intrinsic dimensionality estimators. IEEE Trans PAMI 17:81–86
Zou H (2006) The adaptive Lasso and its oracle properties. J Am Stat Assoc 101:1418–1429
Zou H, Li R (2008) One-step sparse estimates in nonconcave penalized likelihood models (with discussion). Ann Stat 36:1509–1533

Maximum Likelihood Estimation of Intrinsic Dimension

Elizaveta Levina
Department of Statistics
University of Michigan
Ann Arbor MI 48109-1092
elevina@umich.edu

Peter J. Bickel
Department of Statistics
University of California
Berkeley CA 94720-3860
bickel@stat.berkeley.edu

Abstract

We propose a new method for estimating intrinsic dimension of a dataset derived by applying the principle of maximum likelihood to the distances between close neighbors. We derive the estimator by a Poisson process approximation, assess its bias and variance theoretically and by simulations, and apply it to a number of simulated and real datasets. We also show it has the best overall performance compared with two other intrinsic dimension estimators.

1 Introduction

There is a consensus in the high-dimensional data analysis community that the only reason any methods work in very high dimensions is that, in fact, the data are not truly high-dimensional. Rather, they are embedded in a high-dimensional space, but can be efficiently summarized in a space of a much lower dimension, such as a nonlinear manifold. Then one can reduce dimension without losing much information for many types of real-life high-dimensional data, such as images, and avoid many of the "curses of dimensionality". Learning these data manifolds can improve performance in classification and other applications, but if the data structure is complex and nonlinear, dimensionality reduction can be a hard problem.

Traditional methods for dimensionality reduction include principal component analysis (PCA), which only deals with linear projections of the data, and multidimensional scaling (MDS), which aims at preserving pairwise distances and traditionally is used for visualizing data. Recently, there has been a surge of interest in manifold projection methods (Locally Linear Embedding (LLE) [1], Isomap [2], Laplacian and Hessian Eigenmaps [3, 4], and others), which focus on finding a nonlinear low-dimensional embedding of high-dimensional data. So far, these methods have mostly been used for exploratory tasks such as visualization, but they have also been successfully applied to classification problems [5, 6].

The dimension of the embedding is a key parameter for manifold projection methods: if the dimension is too small, important data features are "collapsed" onto the same dimension, and if the dimension is too large, the projections become noisy and, in some cases, unstable. There is no consensus, however, on how this dimension should be determined. LLE [1] and its variants assume the manifold dimension is provided by the user. Isomap [2] provides error curves that can be "eyeballed" to estimate dimension. The charting algorithm, a recent LLE variant [7], uses a heuristic estimate of dimension which is essentially equivalent to the regression estimator of [8] discussed below. Constructing a reliable estimator of intrinsic dimension and understanding its statistical properties will clearly facilitate further applications of manifold projection methods and improve their performance.

We note that for applications such as classification, cross-validation is in principle the simplest solution – just pick the dimension which gives the lowest classification error. However, in practice the computational cost of cross-validating for the dimension is prohibitive, and an estimate of the intrinsic dimension will still be helpful, either to be used directly or to narrow down the range for cross-validation.

In this paper, we present a new estimator of intrinsic dimension, study its statistical properties, and compare it to other estimators on both simulated and real datasets. Section 2 reviews previous work on intrinsic dimension. In Section 3 we derive the estimator and give its approximate asymptotic bias and variance. Section 4 presents results on datasets and compares our estimator to two other estimators of intrinsic dimension. Section 5 concludes with discussion.

2 Previous Work on Intrinsic Dimension Estimation

The existing approaches to estimating the intrinsic dimension can be roughly divided into two groups: eigenvalue or projection methods, and geometric methods. Eigenvalue methods, from the early proposal of [9] to a recent variant [10] are based on a global or local PCA, with intrinsic dimension determined by the number of eigenvalues greater than a given threshold. Global PCA methods fail on nonlinear manifolds, and local methods depend heavily on the precise choice of local regions and thresholds [11]. The eigenvalue methods may be a good tool for exploratory data analysis, where one might plot the eigenvalues and look for a clear-cut boundary, but not for providing reliable estimates of intrinsic dimension.

The geometric methods exploit the intrinsic geometry of the dataset and are most often based on fractal dimensions or nearest neighbor (NN) distances. Perhaps the most popular fractal dimension is the correlation dimension [12, 13]: given a set $S_n = \{x_1, \ldots, x_n\}$ in a metric space, define

$$C_n(r) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbf{1}\{\|x_i - x_j\| < r\}. \tag{1}$$

The correlation dimension is then estimated by plotting $\log C_n(r)$ against $\log r$ and estimating the slope of the linear part [12]. A recent variant [13] proposed plotting this estimate against the true dimension for some simulated data and then using this calibrating curve to estimate the dimension of a new dataset. This requires a different curve for each n, and the choice of calibration data may affect performance. The capacity dimension and packing numbers have also been used [14]. While the fractal methods successfully exploit certain geometric aspects of the data, the statistical properties of these methods have not been studied.

The correlation dimension (1) implicitly uses NN distances, and there are methods that focus on them explicitly. The use of NN distances relies on the following fact: if $X_1, \ldots, X_n$ are an independent identically distributed (i.i.d.) sample from a density $f(x)$ in $\mathbb{R}^m$, and $T_k(x)$ is the Euclidean distance from a fixed point x to its k-th NN in the sample, then

$$\frac{k}{n} \approx f(x)V(m)[T_k(x)]^m, \tag{2}$$

where $V(m) = \pi^{m/2}[\Gamma(m/2+1)]^{-1}$ is the volume of the unit sphere in $\mathbb{R}^m$. That is, the proportion of sample points falling into a ball around x is roughly $f(x)$ times the volume of the ball.

The relationship (2) can be used to estimate the dimension by regressing $\log \bar{T}_k$ on $\log k$ over a suitable range of k, where $\bar{T}_k = n^{-1}\sum_{i=1}^n T_k(X_i)$ is the average of distances from each point to its k-th NN [8, 11]. A comparison of this method to a local eigenvalue method [11] found that the NN method suffered more from underestimating dimension for high-dimensional datasets, but the eigenvalue method was sensitive to noise and parameter settings. A more sophisticated NN approach was recently proposed in [15], where the dimension is estimated from the length of the minimal spanning tree on the geodesic NN distances computed by Isomap.

While there are certainly existing methods available for estimating intrinsic dimension, there are some issues that have not been adequately addressed. The behavior of the estimators as a function of sample size and dimension is not well understood or studied beyond the obvious "curse of dimensionality"; the statistical properties of the estimators, such as bias and variance, have not been looked at (with the exception of [15]); and comparisons between methods are not always presented.

3 A Maximum Likelihood Estimator of Intrinsic Dimension

Here we derive the maximum likelihood estimator (MLE) of the dimension m from i.i.d. observations $X_1, \ldots, X_n$ in $\mathbb{R}^p$. The observations represent an embedding of a lower-dimensional sample, i.e., $X_i = g(Y_i)$, where Y_i are sampled from an unknown smooth density f on $\mathbb{R}^m$, with unknown $m \le p$, and g is a continuous and sufficiently smooth (but not necessarily globally isometric) mapping. This assumption ensures that close neighbors in $\mathbb{R}^m$ are mapped to close neighbors in the embedding.

The basic idea is to fix a point x, assume $f(x) \approx$ const in a small sphere $S_x(R)$ of radius R around x, and treat the observations as a homogeneous Poisson process in $S_x(R)$. Consider the inhomogeneous process $\{N(t,x),\ 0 \le t \le R\}$,

$$N(t,x) = \sum_{i=1}^{n} \mathbf{1}\{X_i \in S_x(t)\} \tag{3}$$

which counts observations within distance t from x. Approximating this binomial (fixed n) process by a Poisson process and suppressing the dependence on x for now, we can write the rate $\lambda(t)$ of the process $N(t)$ as

$$\lambda(t) = f(x)V(m)mt^{m-1} \tag{4}$$

This follows immediately from the Poisson process properties since $V(m)mt^{m-1} = \frac{d}{dt}[V(m)t^m]$ is the surface area of the sphere $S_x(t)$. Letting $\theta = \log f(x)$, we can write the log-likelihood of the observed process $N(t)$ as (see e.g., [16])

$$L(m,\theta) = \int_0^R \log \lambda(t)\, dN(t) - \int_0^R \lambda(t)\, dt$$

This is an exponential family for which MLEs exist with probability $\to 1$ as $n \to \infty$ and are unique. The MLEs must satisfy the likelihood equations

$$\frac{\partial L}{\partial \theta} = \int_0^R dN(t) - \int_0^R \lambda(t)dt = N(R) - e^{\theta}V(m)R^m = 0, \tag{5}$$

$$\begin{aligned} \frac{\partial L}{\partial m} &= \left(\frac{1}{m} + \frac{V'(m)}{V(m)}\right) N(R) + \int_0^R \log t\, dN(t) - \\ &\quad - e^{\theta}V(m)R^m \left(\log R + \frac{V'(m)}{V(m)}\right) = 0. \end{aligned} \tag{6}$$

Substituting (5) into (6) gives the MLE for m:

$$\hat{m}_R(x) = \left[\frac{1}{N(R,x)} \sum_{j=1}^{N(R,x)} \log \frac{R}{T_j(x)} \right]^{-1} . \tag{7}$$

In practice, it may be more convenient to fix the number of neighbors k rather than the radius of the sphere R. Then the estimate in (7) becomes

$$\hat{m}_k(x) = \left[\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{T_k(x)}{T_j(x)} \right]^{-1} . \tag{8}$$

Note that we omit the last (zero) term in the sum in (7). One could divide by $k-2$ rather than $k-1$ to make the estimator asymptotically unbiased, as we show below. Also note that the MLE of θ can be used to obtain an instant estimate of the entropy of f, which was also provided by the method used in [15].

For some applications, one may want to evaluate local dimension estimates at every data point, or average estimated dimensions within data clusters. We will, however, assume that all the data points come from the same "manifold", and therefore average over all observations.

The choice of k clearly affects the estimate. It can be the case that a dataset has different intrinsic dimensions at different scales, e.g., a line with noise added to it can be viewed as either 1-d or 2-d (this is discussed in detail in [14]). In such a case, it is informative to have different estimates at different scales. In general, for our estimator to work well the sphere should be small and contain sufficiently many points, and we have work in progress on choosing such a k automatically. For this paper, though, we simply average over a range of small to moderate values $k = k_1 \ldots k_2$ to get the final estimates

$$\bar{m}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{m}_k(X_i) , \qquad \hat{m} = \frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \bar{m}_k . \tag{9}$$

The choice of k_1 and k_2 and behavior of $\bar{m}_k$ as a function of k are discussed further in Section 4. The only parameters to set for this method are k_1 and k_2, and the computational cost is essentially the cost of finding k_2 nearest neighbors for every point, which has to be done for most manifold projection methods anyway.

3.1 Asymptotic behavior of the estimator for m fixed, $n \to \infty$.

Here we give a sketchy discussion of the asymptotic bias and variance of our estimator, to be elaborated elsewhere. The computations here are under the assumption that m is fixed, $n \to \infty, k \to \infty$, and $k/n \to 0$.

As we remarked, for a given x if $n \to \infty$ and $R \to 0$, the inhomogeneous binomial process $N(t,x)$ in (3) converges weakly to the inhomogeneous Poisson process with rate $\lambda(t)$ given by (4). If we condition on the distance $T_k(x)$ and assume the Poisson approximation is exact, then $\left\{m^{-1} \log(T_k/T_j) : 1 \le j \le k-1\right\}$ are distributed as the order statistics of a sample of size $k-1$ from a standard exponential distribution. Hence $U = m^{-1} \sum_{j=1}^{k-1} \log(T_k/T_j)$ has a Gamma$(k-1,1)$ distribution, and $EU^{-1} = 1/(k-2)$. If we use $k-2$ to normalize, then under these assumptions, to a first order approximation

$$E\left(\hat{m}_k(x)\right) = m, \quad \mathrm{Var}\left(\hat{m}_k(x)\right) = \frac{m^2}{k-3} \tag{10}$$

As this analysis is asymptotic in both k and n, the factor $(k-1)/(k-2)$ makes no difference. There are, of course, higher order terms since $N(t,x)$ is in fact a binomial process with $EN(t,x) = \lambda(t)\left(1 + O(t^2)\right)$, where $O(t^2)$ depends on m.

With approximations (10), we have $E\hat{m} = E\hat{m}_k = m$, but the computation of $\mathrm{Var}(\hat{m})$ is complicated by the dependence among $\hat{m}_k(X_i)$. We have a heuristic argument (omitted for lack of space) that, by dividing $\hat{m}_k(X_i)$ into n/k roughly independent groups of size k each, the variance can be shown to be of order n^{-1}, as it would if the estimators were independent. Our simulations confirm that this approximation is reasonable – for instance, for m-d Gaussians the ratio of the theoretical SD $= C(k_1,k_2)m/\sqrt{n}$ (where $C(k_1,k_2)$ is calculated as if all the terms in (9) were independent) to the actual SD of $\hat{m}$ was between 0.7 and 1.3 for the range of values of m and n considered in Section 4. The bias, however, behaves worse than the asymptotics predict, as we discuss further in Section 5.

4 Numerical Results

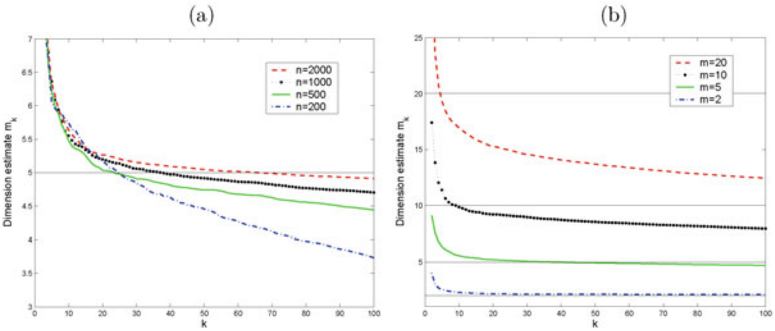


Figure 1: The estimator $\hat{m}_k$ as a function of k. (a) 5-dimensional normal for several sample sizes. (b) Various m-dimensional normals with sample size $n = 1000$.

We first investigate the properties of our estimator in detail by simulations, and then apply it to real datasets. The first issue is the behavior of $\hat{m}_k$ as a function of k. The results shown in Fig. 1 are for m-d Gaussians $N_m(0, I)$, and a similar pattern holds for observations in a unit cube, on a hypersphere, and on the popular "Swiss roll" manifold. Fig. 1(a) shows $\hat{m}_k$ for a 5-d Gaussian as a function of k for several sample sizes n. For very small k the approximation does not work yet and $\hat{m}_k$ is unreasonably high, but for k as small as 10, the estimate is near the true value $m = 5$. The estimate shows some negative bias for large k, which decreases with growing sample size n, and, as Fig. 1(b) shows, increases with dimension. Note, however, that it is the intrinsic dimension m rather than the embedding dimension $p \geq m$ that matters; and as our examples below and many examples elsewhere show, the intrinsic dimension for real data is frequently low.

The plots in Fig. 1 show that the "ideal" range $k_1 \ldots k_2$ is different for every combination of m and n, but the estimator is fairly stable as a function of k, apart from the first few values. While fine-tuning the range $k_1 \ldots k_2$ for different n is possible and would reduce the bias, for simplicity and reproducibility of our results we fix $k_1 = 10$, $k_2 = 20$ throughout this paper. In this range, the estimates are not affected much by sample size or the positive bias for very small k, at least for the range of m and n under consideration.

Next, we investigate an important and often overlooked issue of what happens when the data are near a manifold as opposed to exactly on a manifold. Fig. 2(a) shows simulation results for a 5-d correlated Gaussian with mean 0, and covariance matrix $[\sigma_{ij}] = [\rho + (1-\rho)\delta_{ij}]$, with $\delta_{ij} = \mathbf{1}\{i = j\}$. As ρ changes from 0 to 1, the dimension changes from 5 (full spherical Gaussian) to 1 (a line in $\mathbb{R}^5$), with intermediate values of ρ providing noisy versions.

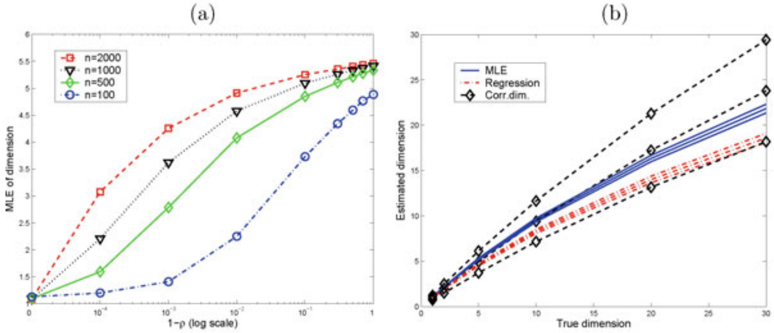


Figure 2: (a) Data near a manifold: estimated dimension for correlated 5-d normal as a function of $1-\rho$. (b) The MLE, regression, and correlation dimension for uniform distributions on spheres with $n = 1000$. The three lines for each method show the mean ± 2 SD (95% confidence intervals) over 1000 replications.

The plots in Fig. 2(a) show that the MLE of dimension does not drop unless ρ is very close to 1, so the estimate is not affected by whether the data cloud is spherical or elongated. For ρ close to 1, when the dimension really drops, the estimate depends significantly on the sample size, which is to be expected: $n = 100$ highly correlated points look like a line, but $n = 2000$ points fill out the space around the line. This highlights the fundamental dependence of intrinsic dimension on the neighborhood scale, particularly when the data may be observed with noise. The MLE of dimension, while reflecting this dependence, behaves reasonably and robustly as a function of both ρ and n.

A comparison of the MLE, the regression estimator (regressing $\log \overline{T}_k$ on $\log k$), and the correlation dimension is shown in Fig. 2(b). The comparison is shown on uniformly distributed points on the surface of an m−dimensional sphere, but a similar pattern held in all our simulations. The regression range was held at $k = 10 \ldots 20$ (the same as the MLE) for fair comparison, and the regression for correlation dimension was based on the first $10 \ldots 100$ distinct values of $\log C_n(r)$, to reflect the fact there are many more points for the $\log C_n(r)$ regression than for the $\log \overline{T}_k$ regression. We found in general that the correlation dimension graph can have more than one linear part, and is more sensitive to the choice of range than either the MLE or the regression estimator, but we tried to set the parameters for all methods in a way that does not give an unfair advantage to any and is easily reproducible.

The comparison shows that, while all methods suffer from negative bias for higher dimensions, the correlation dimension has the smallest bias, with the MLE coming in close second. However, the variance of correlation dimension is much higher than that of the MLE (the SD is at least 10 times higher for *all* dimensions). The regression estimator, on the other hand, has relatively low variance (though always higher than the MLE) but the largest negative bias. On the balance of bias and variance, MLE is clearly the best choice.

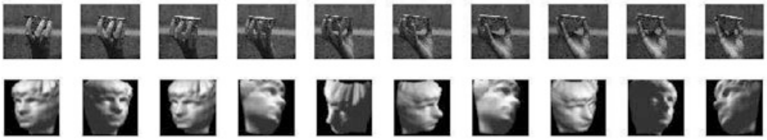


Figure 3: Two image datasets: hand rotation and Isomap faces (example images).

Table 1: Estimated dimensions for popular manifold datasets. For the Swiss roll, the table gives mean(SD) over 1000 uniform samples.

Dataset	Data dim.	Sample size	MLE	Regression	Corr. dim.
Swiss roll	3	1000	2.1(0.02)	1.8(0.03)	2.0(0.24)
Faces	64×64	698	4.3	4.0	3.5
Hands	480×512	481	3.1	2.5	3.9[1]

Finally, we compare the estimators on three popular manifold datasets (Table 1): the Swiss roll, and two image datasets shown on Fig. 3: the Isomap face database[2], and the hand rotation sequence[3] used in [14]. For the Swiss roll, the MLE again provides the best combination of bias and variance.

The face database consists of images of an artificial face under three changing conditions: illumination, and vertical and horizontal orientation. Hence the intrinsic dimension of the dataset should be 3, but only if we had the full 3-d images of the face. All we have, however, are 2-d projections of the face, and it is clear that one needs more than one "basis" image to represent different poses (from casual inspection, front view and profile seem sufficient). The estimated dimension of about 4 is therefore very reasonable.

The hand image data is a real video sequence of a hand rotating along a 1-d curve in space, but again several basis 2-d images are needed to represent different poses (in this case, front, back, and profile seem sufficient). The estimated dimension around 3 therefore seems reasonable. We note that the correlation dimension provides two completely different answers for this dataset, depending on which linear part of the curve is used; this is further evidence of its high variance, which makes it a less reliable estimate that the MLE.

5 Discussion

In this paper, we have derived a maximum likelihood estimator of intrinsic dimension and some asymptotic approximations to its bias and variance. We have shown

[1] This estimate is obtained from the range 500...1000. For this dataset, the correlation dimension curve has two distinct linear parts, with the first part over the range we would normally use, 10...100, producing dimension 19.7, which is clearly unreasonable.

[2] `http://isomap.stanford.edu/datasets.html`

[3] `http://vasc.ri.cmu.edu//idb/html/motion/hand/index.html`

that the MLE produces good results on a range of simulated and real datasets and outperforms two other dimension estimators. It does, however, suffer from a negative bias for high dimensions, which is a problem shared by all dimension estimators. One reason for this is that our approximation is based on sufficiently many observations falling into a small sphere, and that requires very large sample sizes in high dimensions (we shall elaborate and quantify this further elsewhere). For some datasets, such as points in a unit cube, there is also the issue of edge effects, which generally become more severe in high dimensions. One can potentially reduce the negative bias by removing the edge points by some criterion, but we found that the edge effects are small compared to the sample size problem, and we have been unable to achieve significant improvement in this manner. Another option used by [13] is calibration on simulated datasets with known dimension, but since the bias depends on the sampling distribution, and a different curve would be needed for every sample size, calibration does not solve the problem either. One should keep in mind, however, that for most interesting applications intrinsic dimension will not be very high – otherwise there is not much benefit in dimensionality reduction; hence in practice the MLE will provide a good estimate of dimension most of the time.

References

[1] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[2] J. B. Tenenbaum, V. de Silva, and J. C. Landford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in NIPS*, volume 14. MIT Press, 2002.

[4] D. L. Donoho and C. Grimes. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. Technical Report TR 2003-08, Department of Statistics, Stanford University, 2003.

[5] M. Belkin and P. Niyogi. Using manifold structure for partially labelled classification. In *Advances in NIPS*, volume 15. MIT Press, 2003.

[6] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas. Non-linear dimensionality reduction techniques for classification and visualization. In *Proceedings of 8th SIGKDD*, pages 645–651. Edmonton, Canada, 2002.

[7] M. Brand. Charting a manifold. In *Advances in NIPS*, volume 14. MIT Press, 2002.

[8] K.W. Pettis, T.A. Bailey, A.K. Jain, and R.C. Dubes. An intrinsic dimensionality estimator from near-neighbor information. *IEEE Trans. on PAMI*, 1:25–37, 1979.

[9] K. Fukunaga and D.R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Trans. on Computers*, C-20:176–183, 1971.

[10] J. Bruske and G. Sommer. Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE Trans. on PAMI*, 20(5):572–575, 1998.

[11] P. Verveer and R. Duin. An evaluation of intrinsic dimensionality estimators. *IEEE Trans. on PAMI*, 17(1):81–86, 1995.

[12] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica*, D9:189–208, 1983.

[13] F. Camastra and A. Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based approach. *IEEE Trans. on PAMI*, 24(10):1404–1407, 2002.

[14] B. Kegl. Intrinsic dimension estimation using packing numbers. In *Advances in NIPS*, volume 14. MIT Press, 2002.

[15] J. Costa and A. O. Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Trans. on Signal Processing*, 2004. To appear.

[16] D. L. Snyder. *Random Point Processes*. Wiley, New York, 1975.

Journal of Machine Learning Research 7 (2006) 705–732 Submitted 7/05; Revised 12/05; Published 5/06

Some Theory for Generalized Boosting Algorithms

Peter J. Bickel BICKEL@STAT.BERKELEY.EDU
Department of Statistics
University of California at Berkeley
Berkeley, CA 94720, USA

Ya'acov Ritov (corresponding author) YAACOV.RITOV@HUJI.AC.IL
Department of Statistics and The Interdisciplinary Center for Neural Computation
The Hebrew University of Jerusalem
91905 Jerusalem, Israel

Alon Zakai ALONZAKA@POB.HUJI.AC.IL
The Interdisciplinary Center for Neural Computation
The Hebrew University of Jerusalem
91904 Jerusalem, Israel

Editor: Bin Yu

Abstract

We give a review of various aspects of boosting, clarifying the issues through a few simple results, and relate our work and that of others to the minimax paradigm of statistics. We consider the population version of the boosting algorithm and prove its convergence to the Bayes classifier as a corollary of a general result about Gauss-Southwell optimization in Hilbert space. We then investigate the algorithmic convergence of the sample version, and give bounds to the time until perfect separation of the sample. We conclude by some results on the statistical optimality of the L_2 boosting.

Keywords: classification, Gauss-Southwell algorithm, AdaBoost, cross-validation, non-parametric convergence rate

1. Introduction

We consider a standard classification problem: Let $(X,Y),(X_1,Y_1),\ldots,(X_n,Y_n)$ be an i.i.d. sample, where $Y_i \in \{-1,1\}$ and $X_i \in \mathcal{X}$. The goal is to find a good classification rule, $\mathcal{X} \to \{-1,1\}$.

The AdaBoost algorithm was originally defined, Schapire (1990), Freund (1995), and Freund and Schapire (1996) as an algorithm to construct a good classifier by a "weighted majority vote" of simple classifiers. To be more exact, let $\mathcal{H}$ be a set of simple classifiers. The AdaBoost classifier is given by $\operatorname{sgn}\big(\sum_{m=1}^{M} \lambda_m h_m(x)\big)$, where $\lambda_m \in \mathbb{R}$, $h_m \in \mathcal{H}$, are found sequentially by the following algorithm:

0. Let $c_1 = c_2 = \cdots = c_n = 1$, and set $m = 1$.

1. Find $h_m = \arg\min_{h\in\mathcal{H}} \sum_{i=1}^{n} c_i h(X_i)Y_i$. Set

$$\lambda_m = \frac{1}{2}\log\Big(\frac{\sum_{i=1}^{n} c_i + \sum_{i=1}^{n} c_i h_m(X_i)Y_i}{\sum_{i=1}^{n} c_i - \sum_{i=1}^{n} c_i h_m(X_i)Y_i}\Big) = \frac{1}{2}\log\Big(\frac{\sum_{h_m(X_i)=Y_i} c_i}{\sum_{h_m(X_i)\neq Y_i} c_i}\Big).$$

2. Set $c_i \leftarrow c_i \exp\big(-\lambda_m h_m(X_i)Y_i\big)$, and $m \leftarrow m+1$, If $m \le M$, return to step 1.

M is unspecified and can be arbitrarily large.

The success of these methods on many data sets and their "resistance to overfitting"—the test set error continues to decrease even after all the training set observations were classified correctly, has led to intensive investigation to which this paper contributes.

Let $\mathcal{F}_\infty$ be the linear span of $\mathcal{H}$. That is,

$$\mathcal{F}_\infty = \bigcup_{k=1}^{\infty} \mathcal{F}_k, \text{ where } \mathcal{F}_k = \Big\{ \sum_{j=1}^{k} \lambda_j h_j : \lambda_j \in \mathbb{R},\ h_j \in \mathcal{H},\ 1 \le j \le k \Big\} .$$

A number of workers have noted, Breiman (1998,1999), Friedman, Hastie and Tibshirani (2000), Mason, Bartlett, Baxter and Frean (2000), and Schapire and Singer (1999), that the AdaBoost classifier can be viewed as $\operatorname{sgn}\big(F(X)\big)$, where F is found by a greedy algorithm minimizing

$$n^{-1} \sum_{i=1}^{n} \exp\big(-Y_i F(X_i)\big)$$

over $\mathcal{F}_\infty$.

¿From this point of view, the algorithm appeared to be justifiable, since as was noted in Breiman (1999) and Friedman, Hastie, and Tibshirani (2000), the corresponding expression $E\exp\big(-YF(X)\big)$, obtained by replacing the sum by expectation, is minimized by

$$F(X) = \frac{1}{2}\log\Big(P(Y=1|X)/P(Y=-1|X)\Big),$$

provided the linear span $\mathcal{F}_\infty$ is dense in the space $\mathcal{F}$ of all functions in a suitable way. However, it was also noted that the empirical optimization problem necessarily led to rules which would classify every training set observation correctly and hence not approach the Bayes rule whatever be n, except in very special cases. Jiang (2003) established that, for observation centered stumps, the algorithm converged to nearest neighbor classification, a good but rarely optimal rule.

In another direction, the class of objective functions $W(\cdot)$ that can be considered was extended by Friedman, Hastie, and Tibshirani (2000) to other W, in particular, $W(t) = \log(1+e^{-2t})$, whose empirical version they identified with logistic regression in statistics, and $W(t) = -2t + t^2$, which they referred to as "L_2 Boosting" and has been studied, under the name "matching pursuit", in the signal processing community. For all these objective functions, the population optimization of $EW\big(YF(X)\big)$ over $\mathcal{F}$ leads to a solution such that $\operatorname{sgn}F(X)$ is the Bayes rule. Friedman et al. also introduced consideration of other algorithms for the empirical optimization problem. Lugosi and Vayatis (2004) added regularization, changing the function whose expectation (both empirically and in the population) is to be minimized from $W\big(YF(X)\big)$ to $W_n\big(YF(X)\big)$ where $W_n \to W$ as $n \to \infty$. Bühlmann and Yu (2003) considered L_2 boosting starting from very smooth functions. We shall elaborate on this later.

We consider the behavior of the algorithm as applied to the sample $(Y_1,X_1),\ldots,(Y_n,X_n)$, as well to the "population", that is when means are replaced by expectations and sums by probabilities. The structure of, and the differences between, the population and sample versions of the optimization problem has been explored in various ways by Jiang (2003), Zhang and Yu (2003), Bühlmann (2003), Bartlett, Jordan, and McAuliffe (2003), Bickel and Ritov (2003).

Our goal in this paper is

1. To clarify the issues through a few simple results.
2. To relate our work and that of Bühlmann (2003), Bühlmann and Yu (2003), Lugosi and Vayatis (2004), Zhang (2004), Zhang and Yu (2003) and Bartlett, Jordan, and McAuliffe (2003) to the minimax results of Mammen and Tsybakov (1999), Baraud (2001) and Tsybakov (2001).

In Section 2 we will discuss the population version of the basic boosting algorithms and show how their convergence and that of more general greedy algorithms can be derived from a generalization of Theorem 3 of Mallat and Zhang (1993) with a simple proof. The result can, we believe, also be derived from the even more general theorem of Zhang and Yu (2003), but our method is simpler and the results are transparent.

In Section 3 we show how Bayes consistency of various sample algorithms when suitably stopped or of sample algorithms based on minimization of a regularized W follow readily from population convergence of the algorithms and indicate how test bed validation can be used to do this in a way leading to optimal rates (in Section 4).

In Section 5 we address the issue of bounding the time to perfect separation of the different boosting algorithm (including the standard AdaBoost).

Finally in Section 6 we show how minimax rate results for estimating $E(Y|X)$ may be attained for a "sieve" version of the L_2 boosting algorithm, and relate these to results of Baraud (2001), Lugosi and Vayatis (2004), Bühlmann and Yu (2003), Barron, Birgé, Massart(1999) and Bartlett, Jordan and McAuliffe (2003). We also discuss the relation of these results to classification theory.

2. Boosting "Population" Theorem

We begin with a general theorem on Gauss-Southwell optimization in vector space. It is, in part, a generalization of Theorem 1 of Mallat and Zhang (1993) with a simpler proof. A second part relates to procedures in which the step size is regularized cf. Zhang and Yu (2003) and Bartlett et al. (2003). We make the boosting connection after its statement.

Let w be a real, bounded from below, convex function on a vector space $\mathbb{H}$. Let $\mathcal{H} = \mathcal{H}' \cup (-\mathcal{H}')$, where $\mathcal{H}'$ is a subset of $\mathbb{H}$ whose members are linearly independent, with linear span $\mathcal{F}_\infty = \{\sum_{m=1}^{k} \lambda_m h_m : \lambda_j \in \mathbb{R},\ h_j \in \mathcal{H},\ 1 \le j \le k,\ 1 \le k < \infty\}$. We assume that $\mathcal{F}_\infty$ is dense in $\mathbb{H}$, at least in the sense that $\{w(f) : f \in \mathcal{F}_\infty\}$ is dense in the image of w. We define two relaxed Gauss-Southwell "algorithms".

Algorithm I: For $\alpha \in (0,1]$, and given $f_1 \in \mathbb{H}$, find inductively $f_2, f_3, \ldots, \ldots$ by, $f_{m+1} = f_m + \lambda_m h_m$, $\lambda_m \in \mathbb{R}$, $h_m \in \mathcal{H}$ and

$$w(f_m + \lambda_m h_m) \le \alpha \min_{\lambda \in \mathbb{R},\, h \in \mathcal{H}} w(f_m + \lambda h) + (1-\alpha) w(f_m)\ . \tag{1}$$

Generalize Algorithm I to :

Algorithm II: Like Algorithm I, but replace (1) by

$$w(f_m + \lambda_m h) + \gamma \lambda_m^2 \le \alpha \min_{\lambda \in \mathbb{R},\, h \in \mathcal{H}} (w(f_m + \lambda h) + \gamma \lambda^2) + (1-\alpha) w(f_m)\ .$$

There are not algorithms in the usual sense since they do not specify a unique sequence of iterations but our theorems will apply to any sequence generated in this way. Technically, this scheme is used in the proof of Theorem 3. The standard boosting algorithms theoretically correspond to $\alpha = 1$, although in practice, since numerical minimization is used, α may equal 1 only approximately. Our generalization makes for a simple proof and covers the possibility that the minimum of $w(f_m + \lambda h)$ over $\mathcal{H}$ and $\mathbb{R}$ is not assumed, or multiply assumed. Let $\omega_0 = \inf_{f \in \mathcal{F}_\infty} w(f) > -\infty$. Let $w'(f;h)$ the linear operator of the Gataux derivative at $f \in \mathcal{F}_\infty$ in the direction $h \in \mathcal{F}_\infty$: $w'(f;h) = \partial w(f+\lambda h)/\partial\lambda\big|_{\lambda=0}$, and let $w''(f;h)$ be the second derivative of w at f in the direction h: $w''(f,h) \equiv \partial^2 w(f+\lambda h)/\partial\lambda^2\big|_{\lambda=0}$ (both derivative are assumed to exist). We consider the following conditions.

GS1. For any c_1 and c_2 such that $\omega_0 < c_1 < c_2 < \infty$,

$$\begin{aligned} 0 < \ &\inf\{w''(f,h) : c_1 < w(f) < c_2,\ h \in \mathcal{H}\} \\ &\le \sup\{w''(f,h) : w(f) < c_2,\ h \in \mathcal{H}\} < \infty. \end{aligned}$$

GS2. For any $c_2 < \infty$,

$$\sup\{w''(f,h) :\ w(f) < c_2,\ h \in \mathcal{H}\} < \infty\,.$$

Theorem 1 *Under Assumption* GS1*, any sequence of functions generated according to Algorithm I satisfies:*

$$w(f_m) \le \omega_0 + c_m$$

and if $c_m > 0$*:*

$$w(f_m) - w(f_{m+1}) \ge \xi(w(f_m)) > 0$$

where the sequence $c_m \to 0$ *and the function* $\xi(\cdot)$ *depend only on* α*, the initial points of the iterates, and* $\mathcal{H}$*. The same conclusion holds under Condition* GS2 *for any sequence* f_m *generated according to algorithm II.*

The proof can be found in Appendix A.

Remark:

1. Condition GS2 of Theorem 1 guarantees that $\sum_{m=1}^{\infty} \lambda_m^2 < \infty$. It can be replaced by any other condition that guarantees the same, for example, limiting the step size, replacing the penalty by other penalties, etc.

2. It will be clear from the proof in Appendix A that if w'' is bounded away from 0 and ∞ then c_m is of order $(\log m)^{-\frac{1}{2}}$ so that we, in fact, have an approximation rate – but it is so slow as to be essentially useless. On the other hand, with strong conditions such as orthonormality of the elements of $\mathcal{H}$, and $\mathcal{H}$ a classical approximation class such as trigonometric functions we expect, with L_2 boosting, to obtain rates such as $m^{-1/2}$ or better.

Let $(X,Y) \sim P, X \in \mathcal{X}, Y \in \{-1,1\}$. Let $\mathcal{H} \subset \{h : \mathcal{X} \to [-1,1]\}$ be a symmetric set of functions. In particular, $\mathcal{H}$ can, but need not, be a set of classifiers such as trees with

$$\mathcal{H} = -\mathcal{H}\,. \tag{2}$$

Given a loss function $W:\mathbb{R}\to\mathbb{R}^+$, we consider a greedy sequential procedure for finding a function F that minimizes $EW\big(YF(X)\big)$. That is, given $F_0\in\mathcal{H}$ fixed, we define for $m\geq 0$:

$$\begin{aligned}
\lambda_m(h) &= \arg\min_{\lambda\in\mathbb{R}} EW\Big(Y\big(F_m(X)+\lambda h(X)\big)\Big)\\
h_m &= \arg\min_{h\in\mathcal{H}} EW\Big(Y\big(F_m(X)+\lambda_m(h)h(X)\big)\Big)\\
F_{m+1} &= F_m+\lambda_m(h_m)h_m.
\end{aligned}$$

Assume, wlog (without loss of generality), by shifting and rescaling, that $W(0)=-W'(0)=1$. Note that by Bartlett et al. (2003), $W'(0)<0$ is necessary and sufficient for population consistency defined below. We can_suppose again wlog in view of (2), that $\lambda_m\geq 0$. Define $\mathcal{F}_k$ and $\mathcal{F}_\infty$ as in Section 1 and let $\mathcal{F}\equiv\mathcal{F}_\infty$ be the closure of $\mathcal{F}_\infty$ in convergence in probability:

$$\begin{aligned}
\mathcal{F} &\equiv \big\{F:\exists F_m\in\mathcal{F}_m,\ F_m(X)\xrightarrow{P}F(X)\big\}\\
F_\infty &\equiv \arg\min_{F\in\mathcal{F}}\ EW\big(YF(X)\big)
\end{aligned}$$

If $\operatorname{sgn}F_\infty$ is the Bayes rule for 0-1 loss, we say that F_∞ is population consistent for classification, "calibrated" in the Bartlett et al. terminology. Let

$$\begin{aligned}
p(X) &\equiv P(Y=1|X)\\
\widetilde{W}(x,d) &\equiv p(x)W\big(d\big)+\big(1-p(x)\big)W\big(-d\big).\\
\widetilde{W}(F) &\equiv \widetilde{W}(X,F(X))
\end{aligned}$$

By the assumptions below F_∞ is the unique function such that $\widetilde{W}'(F_\infty)=0$ with probability 1, where $\widetilde{W}'(F)=\widetilde{W}'(X,F(X))$ and $\widetilde{W}'(x,d)=\partial W(x,d)/\partial d$. Define $\widetilde{W}''$ similarly.

Here are some conditions.

P1. $P[p(X)=0 \text{ or } 1]=0$.

P2. W is twice differentiable and convex on $\mathbb{R}$.

P3. $\mathcal{H}$ is closed and compact in the weak topology. $\mathcal{F}$ is the set of all measurable functions on $\mathcal{X}$.

P4. $\widetilde{W}''(F)$ is bounded above and below on $\{F:\ c_1<\widetilde{W}(F)<c_2\}$ for all c_1,c_2 such that

$$\inf_{F\in\mathcal{F}} E\widetilde{W}(F)<c_1<c_2<E\widetilde{W}(F_0).$$

P5. $F_\infty\in L_2(P)$.

Note that P1 and P2 imply that $\widetilde{W}(x,d)\to\infty$ as $|d|\to\infty$, which ensures that F_∞ is finite almost anywhere. Condition P1, which says that no point can be classified with absolute certainty, is only needed technically to ensure that $\widetilde{W}(x,d)\to\infty$ as $|d|\to\infty$, even if W itself is monotone. It is not needed for L_2 boosting.

Conditions P2 and P4 ensure that along the optimizing path W behaves locally like $W_0(t)=-2t+t^2$ corresponding to L_2 boosting. They are more stringent than we would like and, in particular, rule out W such as the "hinge" appearing in SVM. More elaborate arguments such as those of Zhang and Yu (2003) and Bartlett et al. (2003) can give somewhat better results.

The functions commonly appearing in boosting such as, $W_1(t) = e^{-t}$, $W_2(t) = -2t + t^2$, $W_3(t) = -\log(1 + e^{-2t})$ satisfy condition P4 if P1 also holds. This is obvious for W_2. For W_1 and W_3, it is clear that P4 holds, if P1 does, since otherwise $E\widetilde{W}\big(YF_m(X)\big) \to \infty$. The conclusions of Theorem 2 continue to hold if $h \in \mathcal{H} \implies |h| \geq \delta > 0$ since then below $w''(F;h) = Eh^2(X)\widetilde{W}\big(F(X)\big) \geq \delta^2 E\widetilde{W}(F(X))$ and P4 follows. Note that if $|h| \not\equiv 1$ the λ optimization step requires multiplying λ^2 by $Eh^2(x)$.

We have,

Theorem 2 *If $\mathcal{H}$ is a set of classifiers,* $(h^2 \equiv 1)$ *and Assumptions* P2 – P5 *hold, then*

$$F_m(X) \xrightarrow{P} F_\infty(X) ,$$

and the misclassification error, $P\big(YF_m(X) \leq 0\big) \to P[YF_\infty(X) \leq 0]$, *the Bayes risk.*

Proof Identify $w(F) = EW\big(YF(X)\big) = E\widetilde{W}\big(F(X)\big)$. Then,

$$w''(F,h) = Eh^2(X)\widetilde{W}''\big(F(X)\big) = E\widetilde{W}''\big(F(X)\big)$$

and (P4) can be identified with condition GS1 of Theorem 1. Thus,

$$E\widetilde{W}\big(F_m(X)\big) \to E\widetilde{W}(F_\infty(X)) .$$

Since,

$$E\widetilde{W}\big(F_m(X)\big) - E\widetilde{W}\big(F_\infty(X)\big) = E\Big((F_\infty - F_m)^2 \int_0^1 \widetilde{W}''\big((1-\lambda)(X)F_\infty(X) + \lambda F_m(X)\big)\lambda d\lambda\Big) \to 0 ,$$

the conclusion of Theorem 2 follows from (P4). The second assertion is immediate. ■

3. Consistency of the Boosting Algorithm

In this section we study the Bayes consistency properties of the sample versions of the boosting algorithms we considered in Section 2. In particular, we shall

(i) Show that under mild additional conditions, there will exist a random sequence $m_n \to \infty$ such that $\hat{F}_{m_n} \xrightarrow{P} F_\infty$, where $\hat{F}_m$ is defined below as the mth sample iterate, and moreover, that such a sequence can be determined using the data.

(ii) Comment on the relationship of this result to optimization for penalized versions of W. The difference is that the penalty forces $m < \infty$ to be optimal while with us, cross-validation (or a test bed sample) determines the stopping point. We shall see that the same dichotomy applies later, when we "boost" using the method of sieves for nonparametric regression studied by Barron, Birge and Massart (1999) and Baraud (2001).

3.1 The Golden Chain Argument

Here is a very general framework. This section is largely based on Bickel and Ritov (2003).

Let $\Theta_1 \subset \Theta_2 \subset \dots$ be a sequence of sets contained in a separable metric space, $\Theta = \overline{\cup \Theta_m}$ where $\bar{}$ denotes closure. Let $\Pi_m : \Theta_m \to 2^{\Theta_{m+1}}$ be a sequence of point to set mappings. Let K be a target function, and $\vartheta_\infty = \arg\min_{\vartheta\in\Theta} K(\vartheta)$. Finally, let $\hat{K}_n$ be a sample based approximation of K. We assume:

G1. $K : \Theta \to \mathbb{R}$ is strictly convex, with a unique minimizer ϑ_∞.

Our result is applicable to loosely defined algorithms. In particular we want to be able to consider the result of the algorithm applied to the data as if it were generated by a random algorithm applied to the population. We need therefore, the following definitions. Let $\mathcal{S}(\vartheta_0, \alpha)$ be the set of all sequences $\bar{\vartheta}_m \in \Theta_m$, $m = 0, 1, \dots$ with $\bar{\vartheta}_0 = \vartheta_0$ and satisfying:

$$
\begin{aligned}
&\bar{\vartheta}_{m+1} \in \Pi_m(\bar{\vartheta}_m)\\
&K(\bar{\vartheta}_{m+1}) \le \alpha \inf_{\vartheta\in\Pi_m(\bar{\vartheta}_m)} K(\vartheta) + (1-\alpha) K(\bar{\vartheta}_m).
\end{aligned}
$$

The resemblance to Gauss-Southwell Algorithm I and the boosting procedures is not accidental. Suppose the following uniform convergence criterion is satisifed:

G2. If $\{\bar{\vartheta}_m\} \in \mathcal{S}(\vartheta_0, \alpha)$ with any initial ϑ_0, then $K(\bar{\vartheta}_m) - K(\bar{\vartheta}_{m+1}) \ge \xi\big(K(\bar{\vartheta}_m) - K(\vartheta_\infty)\big)$, for $\xi(\cdot) > 0$ strictly increasing, and $K(\bar{\vartheta}_m) - K(\vartheta_\infty) \le c_m$ where $c_m \to 0$ uniformly over $\mathcal{S}(\vartheta_0, \alpha)$.

In boosting, given P, $\Theta = \{F(X), F \in \tilde{\mathcal{F}}\}$ with a metric of convergence in probability, $\Theta_m = \{\sum_{j=1}^m \lambda_j h_j, h_j \in \mathcal{H}\}$, $\Pi_m(F) = \Pi(F) = \{F + \lambda h, \lambda \in \mathbb{R}, h \in \mathcal{H}\}$, and $K(F) = \mathrm{E} W\big(YF(X)\big)$. Condition G2, follows from the conclusion of Theorem 1.

Now suppose $\hat{K}_n(\cdot)$ is a sequence of random functions on Θ, empirical entities that resemble the population K. Let $\hat{\mathcal{S}}_n(\vartheta_0, \alpha')$ be the set of all sequences $\hat{\vartheta}_{0,n}, \hat{\vartheta}_{1,n} \dots$, such that $\hat{\vartheta}_{0,n} = \vartheta_0$, and

$$
\begin{aligned}
&\hat{\vartheta}_{m+1,n} \in \Pi_m(\hat{\vartheta}_{m,n})\\
&\hat{K}_n(\hat{\vartheta}_{m+1,n}) \le \alpha' \min\{\hat{K}_n(\vartheta) :\ \vartheta \in \Pi_m(\hat{\vartheta}_{m,n})\} + (1-\alpha')\hat{K}_n(\hat{\vartheta}_{m,n}).
\end{aligned}
$$

We assume

G3. $\hat{K}_n$ is convex, and for all integer m, $\sup\{|\hat{K}_n(\vartheta) - K(\vartheta)| : \vartheta \in A_m\} \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$, for a sequence $A_m \subset \Theta_m$ such that $P(\hat{\vartheta}_{m,n} \in A_m) \to 1$.

In boosting, $\hat{K}_n(F) = n^{-1}\sum_{i=1}^n W\big(Y_i F(X_i)\big)$, $K(F) = E_p\big(YF(X)\big)$

The sequence $\{\bar{\vartheta}_m\}$ is the golden chain we try to follow using the obscure information in the sample.

We now state and prove,

Theorem 3 *If assumptions* G1– G3 *hold, and* $\alpha' \in (0,1]$, *then for any sequence* $\{\hat{\vartheta}_{m,n}\} \in \hat{\mathcal{S}}(\vartheta_0, \alpha')$, *there exists a subsequence* $\{\hat{m}_n\}$ *such that* $K(\hat{\vartheta}_{\hat{m}_n,n}) \xrightarrow{\text{P}} K(\vartheta_\infty)$.

Proof

Fix ϑ_0 and α, $\alpha < \alpha'$. Let $M_n \to \infty$ be some sequence, and let $\hat{m}_n = \arg\min_{m \le M_n} K(\hat{\vartheta}_{m,n})$. We need to prove that $K(\hat{\vartheta}_{\hat{m}_n,n}) \xrightarrow{\text{p}} K(\vartheta_\infty)$. We will prove this by contradiction. Suppose otherwise:

$$\inf_{m \le M_n} K(\hat{\vartheta}_{m,n}) - K(\vartheta_\infty) \ge c_1 > 0, \quad n \in \mathcal{N} \tag{3}$$

where $\mathcal{N}$ is unbounded with positive probability. Let $\varepsilon_{m,n} \equiv \sup_{\vartheta \in A_m} |K(\vartheta) - \hat{K}_n(\vartheta)|$. For any fixed m, $\varepsilon_{m,n} \xrightarrow{\text{a.s.}} 0$ by G3. Let

$$m_n = \arg\max\Big\{m' \le M_n : \forall m \le m', \varepsilon_{m-1,n} + 2\varepsilon_{m,n} < (\alpha' - \alpha)\xi(c_1) \;\&\; \hat{\vartheta}_{m,n} \in A_m\Big\}.$$

Clearly, $m_n \xrightarrow{\text{p}} \infty$, and for any $m \le m_n$, assuming (3):

$$\begin{aligned}
K(\hat{\vartheta}_{m,n}) &\le \hat{K}_n(\hat{\vartheta}_{m,n}) + \varepsilon_{m,n} \\
&\le \alpha' \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} \hat{K}_n(\vartheta) + (1-\alpha')\hat{K}_n(\hat{\vartheta}_{m-1,n}) + \varepsilon_{m,n} \\
&\le \alpha' \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta) + (1-\alpha')K(\hat{\vartheta}_{m-1,n}) + \varepsilon_{m-1,n} + 2\varepsilon_{m,n} \\
&= \alpha \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta) + (1-\alpha)K(\hat{\vartheta}_{m-1,n}) \\
&\qquad - (\alpha' - \alpha)\big(K(\hat{\vartheta}_{m-1,n}) - \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta)\big) + \varepsilon_{m-1,n} + 2\varepsilon_{m,n} \\
&\le \alpha \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta) + (1-\alpha)K(\hat{\vartheta}_{m-1,n}) \\
&\qquad - (\alpha' - \alpha)\xi\Big(K(\hat{\vartheta}_{m,n}) - K(\vartheta_\infty)\Big) + \varepsilon_{m-1,n} + 2\varepsilon_{m,n} \\
&\le \alpha \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta) + (1-\alpha)K(\hat{\vartheta}_{m-1,n}) \\
&\qquad - (\alpha' - \alpha)\xi\Big(c_1\Big) + \varepsilon_{m-1,n} + 2\varepsilon_{m,n} \\
&\le \alpha \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta) + (1-\alpha)K(\hat{\vartheta}_{m-1,n}) \text{ for all } m \le m_n .
\end{aligned}$$

Thus, there is a sequence $\{\bar{\vartheta}_1^{(n)}, \bar{\vartheta}_2^{(n)}, \dots\} \in \mathcal{S}(\vartheta_0, \alpha)$, such that $\bar{\vartheta}_m^{(n)} = \hat{\vartheta}_{m,n}$, $m \le m_n$. Hence, by Assumption G2, $K(\vartheta_{m_n,n}) \le K(\vartheta_\infty) + c_{m_n}$, where $\{c_m\}$ is independent of n, and $c_m \to 0$. Therefore, since $m_n \to \infty$, $K(\hat{\vartheta}_{m_n,n}) \to K(\vartheta_\infty)$, contradicting (3). ■

In fact we have proved that sequences m_n can be chosen in the following way involving K.

Corollary 4 *Let M_n be* any *sequence tending to ∞. Let $\tilde{m}_n = \arg\min\{K(\hat{\vartheta}_{m,n}) : 1 \le m \le M_n\}$. Then, under* G1 – G3, $\hat{\vartheta}_{\tilde{m}_n} \xrightarrow{P} \vartheta_\infty$.

To find $\hat{\vartheta}_{\hat{m}_n,n}$ which are totally determined by the data determining $\hat{K}_n$, we need to add some information about the speed of convergence of $\hat{K}_n$ to K on the "sample" iterates. Specifically, suppose we can determine, in advance, $M_n^* \to \infty$, $\varepsilon_n \to 0$ such that,

$$P[\sup\{|\hat{K}_n(\hat{\vartheta}_{m,n}) - K(\hat{\vartheta}_{m,n})| : 1 \le m \le M_n^*\} \ge \varepsilon_n] \le \varepsilon_n .$$

Then $\hat{m}_n = \arg\min\{\hat{K}_n(\hat{\vartheta}_{m,n}) :\ 1 \le m \le M_n^*\}$ yields an appropriate $\hat{\vartheta}_{\hat{m}_n}$ sequence. We consider this in Section 4. Before that we return to the application of the result of this section to boosting.

3.2 Back to Boosting

We return to boosting, where we consider $\Theta_m = \{\sum_{j=1}^m \lambda_j h_j : \lambda_j \in \mathbb{R}, h_j \in \mathcal{H}\}$, and therefore $\Pi_m \equiv \Pi$, $\Pi(\vartheta) = \{\vartheta + \lambda h, \lambda \in \mathbb{R}, h \in \mathcal{H}\}$. To simplify notation, for any function $a(X,Y)$, let $P_n a(X,Y) = n^{-1}\sum_{i=1}^n a(X_i,Y_i)$ and $Pa(X,Y) = Ea(X,Y)$. Finally, we identify $\hat{\vartheta}_{m,n} = \sum_{j=1}^m \hat{\lambda}_j \hat{h}_j = \sum_{j=1}^m \hat{\lambda}_{j,n}\hat{h}_{j,n}$.

We assume further

GA1. $W(\cdot)$ is of bounded variation on finite intervals.

GA2. $\mathcal{H}$ has finite L_1 bracketing entropy.

GA3. There are finite $a_1, a_2, \ldots$ such that $\sup_n \sum_{j=1}^m |\hat{\lambda}_{j,n}| \le a_m$ with probability 1.

Theorem 5 *Suppose the conclusion of Theorem 1 and Conditions* GA1–GA3 *are satisfied, then conditions* G2, G3 *are satisfied.*

Proof Condition G2 follows from Theorem 1. It remains to prove the uniform convergence in Condition G3. However, GA2 and GA3 imply that $\mathcal{F} \equiv \{F : F = \sum_{j=1}^m \lambda_j h_j, h_j \in \mathcal{H}, |\lambda_j| \le M\}$ has finite L_1 bracketing entropy. Since W can be written as the difference of two monotone functions $\{W(YF) : F \in \mathcal{F}\}$ inherits this property. The result follows from Bickel and Millar (1991), Proposition 2.1. ■

4. Test Bed Stopping

Again we face the issue of data dependent and in some way optimal selection of $\hat{m}_n$. We claim that this can be achieved over a wide range of possible rates of convergence of $EW\big(\hat{F}_{\hat{m}_n}(YX)\big)$ to $EW\big(F_\infty(YX)\big)$ by using a test bed sample to pick the estimator. The following general result plays a key role.

Let $B = B_n \to \infty$, and let $(X,Y), (X_1,Y_1), \ldots, (X_{n+B}, Y_{n+B})$ be i.i.d. P, $X \in \mathcal{X}$, $|Y| \le 1$. Let $\hat{\vartheta}_m : \mathcal{X} \to \mathbb{R}$, $1 \le m \le m_n$ be data dependent functions which depend only on $(X_1,Y_1), \ldots, (X_n,Y_n)$ which are predictors of Y. For $g, g_1, g_2 :\ \mathcal{X} \times \mathbb{R} \to \mathbb{R}$, given P, define

$$\begin{aligned}
\langle g_1, g_2\rangle_* &\equiv \frac{1}{B_n}\sum_{b=1}^{B_n} g_1(X_{b+n}, Y_{b+n}) g_2(X_{b+n}, Y_{b+n}) \\
\langle g_1, g_2\rangle_P &\equiv P\big(g_1(X,Y)g_2(X,Y)\big) = \int g_1(x,y)g_2(x,y)dP(x,y) \\
\|g\|_*^2 &\equiv \langle g_1, g_2\rangle_* \\
\|g\|_P^2 &\equiv \langle g_1, g_2\rangle_P
\end{aligned}$$

Let,

$$\tau = \arg\min\{\|Y - \hat{\vartheta}_m(X)\|_*^2 :\ 1 \le m \le M_n\}$$

and $\hat{\vartheta}_\tau$ be the selected predictor. Similarly, let

$$o = \arg\min\{\|Y - \hat{\vartheta}_m(X)\|_P^2 : \ 1 \le m \le M_n\}$$

and $\hat{\vartheta}_O$ be the corresponding predictor.

That is, $\hat{\vartheta}_O(X,Y)$ is the predictor an "oracle" knowing P and (X_i, Y_i), $1 \le i \le n$ would pick from $\hat{\vartheta}_1, \dots, \hat{\vartheta}_{M_n}$ to minimize squared error loss. Let $\vartheta_O(X) \equiv E_P(Y|X)$, the Bayes predictor. Let $\mathcal{P}$ be a set of probabilities and $r_n \equiv \sup\{E_P\|\hat{\vartheta}_O - \vartheta_O\|_P^2 : \ P \in \mathcal{P}\}$.

The following result is due to Györfi et al. (2002) (Theorem 7.1), although there it is stated in the form of an oracle inequality. We need the following condition:

C. $B_n r_n / \log M_n \to \infty$.

Theorem 6 *(Györfi et al.) Suppose condition **C** is satisfied, and* $|Y| \le 1$, $\|\hat{\vartheta}_m\|_\infty \le 1$. *Then,*

$$\sup\{\left|E_P(Y - \hat{\vartheta}_\tau)^2 - E_P(Y - \hat{\vartheta}_O)^2\right| : P \in \mathcal{P}\} = o(r_n).$$

Condition **C** very simply asks that the test sample size B_n be large only: (i) In terms of r_n, the minimax rate of convergence; (ii) In terms of the logarithm of the number of procedures being studied. If $|Y| \le 1$, there is no loss in requiring $\|\hat{\vartheta}_m\|_\infty \le 1$, since we could also replace $\hat{\vartheta}_m$ by its truncation at ± 1, minimizing the L_2 cross validated test set risk. Along similar lines, using $\text{sgn}(\hat{\vartheta}_m)$ is equivalent to cross validating the probability of misclassification for these rules, since if $\hat{\vartheta}_m, Y \in \{-1, 1\}$, $E(Y - \hat{\vartheta}_m)^2 = 4P(\hat{\vartheta}_m \ne Y)$.

As we shall see in Section 6, typically $r_n = n^{-1+\delta}$, and M_n is at most polynomial in n. If n/B_n is slowly varying, we can check that the conditions hold. Essentially we can only not deal with r_n of order $n^{-1}\log n$.

5. Algorithmic Speed of Convergence

We consider now the time it takes the sample algorithm to convergence. The fact that the algorithm converges follows from Theorem 1. We show in this section that in fact the algorithm perfectly separates the data (*perfect separation* is achieved when $Y_i F_m(x_i) > 0$ for all $i = 1, \dots, n$) after no more than $c_1 n^2$ steps. Perfect separation is equivalent to empirical misclassification error 0.

The randomness considered in this section comes only from the Y_i, while the design points are considered fixed. We denote them, therefore, by lower case $x_1, \dots, x_n$. We consider the following assumptions:

O1. W has regular growth in the sense that $W'' < \kappa(W + 1)$ for some $\kappa < \infty$. Assume, wlog, that $W(0) = -W'(0) = 1$.

O2. Suppose $x_1, \dots, x_n$ are all different Then the points can be finitely isolated by $\mathcal{H}$ in the sense that there is k and positive $\alpha_1, \dots, \alpha_k$ such that for every i there are $h_1, \dots, h_k \in \mathcal{H}$ such that $\sum_{j=1}^k \alpha_j h_j(x_s) = 1$ if $s = i$, and 0 otherwise. Assume further, as usual, that if $h \in \mathcal{H}$ then $h^2 \equiv 1$ and $-h \in \mathcal{H}$.

Condition O1 is satisfied by all the loss functions mentioned in the introduction. Condition O2 is satisfied, for example by stumps, trees, and any $\mathcal{H}$ whose span includes indicators of small sets with arbitrary location. In particular, if $x_i \in \mathbb{R}$, $x_1 < x_2 < \cdots < x_n$, and $\mathcal{H} = \{\text{sgn}(\cdot - x), x \in \mathbb{R}\}$, we can then take $\alpha_1 = \alpha_2 = 1$, $h_1(\cdot) = \text{sgn}\big(\cdot - (x_{i-1} + x_i)/2\big)$, and $h_2(\cdot) = -\text{sgn}\big(\cdot - (x_i + x_{i+1})/2\big)$

Theorem 7 *Suppose assumptions* O1 *and* O2 *are satisfied and the algorithm starts with* $F_0(0) = 0$. *If* $Y_iF_m(x_i) < 0$ *for at least one i, then*

$$\frac{1}{n}\sum_{i=1}^{n} W\big(Y_iF_m(x_i)\big) - \frac{1}{n}\sum_{i=1}^{n} W\big(Y_iF_{m+1}(x_i)\big) \geq \frac{1}{2\kappa\big(n\sum_{j=1}^{k}\alpha_j\big)^2}.$$

Hence, the boosting algorithm perfectly separates the data after at most $2\kappa(n\sum_{j=1}^{k}|\alpha_j|)^2$ *steps.*

Proof Let, for i such that $Y_iF_m(x_i) < 0$,

$$f_m(\lambda;h) = n^{-1}\sum_{s=1}^{n} W\Big(Y_i\big(F_m(x_s) + \lambda h(x_s)\big)\Big),$$

and $f'_m(0;h) = df_m(\lambda;h)/d\lambda\big|_{\lambda=0}$. Consider $h_1,\dots,h_k$ as in assumption O2. Replace h_j by $-h_j$ if necessary to ensure that $Y_i\sum_{j=1}^{k}\alpha_j h_j(x_s) = \delta_{si}$. Then

$$\begin{aligned}\sum_{j=1}^{k}\alpha_j f'_m(0;h_j) &= n^{-1}\sum_{j=1}^{k}\alpha_j\sum_{s=1}^{n} W'\big(Y_iF_m(x_s)\big)Y_ih_j(x_s)\\ &= n^{-1}W'\big(Y_iF_m(x_i)\big).\end{aligned}$$

Hence

$$\inf_{h\in\mathcal{H}} f'_m(0;h) \leq \frac{1}{n\sum_{j=1}^{k}\alpha_j}\min_i W'(Y_iF_m(x_i)) \leq \frac{W'(0)}{n\sum_{j=1}^{k}\alpha_j} = \frac{-1}{n\sum_{j=1}^{k}\alpha_j}, \tag{4}$$

since $Y_iF_m(x_i) < 0$ for at least one i.

Let $\bar{h}$ be the minimizer of $f'_m(0,h)$. Note that in particular $f'_m(0;\bar{h}) < 0$. The function $f_m(\cdot;\bar{h})$ is convex, hence it is decreasing in some neighborhood of 0. Denote by $\bar{\lambda}$ its minimizer. Consider the Taylor expansion:

$$\begin{aligned}f_m(\bar{\lambda};\bar{h}) &= f_m(0;\bar{h}) + \bar{\lambda}f'_m(0;\bar{h}) + \frac{\bar{\lambda}^2}{2n}\sum_{s=1}^{n} W''\Big(Y_i\big(F_m(x_s) + \tilde{\lambda}(\lambda)\bar{h}(x_s)\big)\Big)\\ &= f_m(0;\bar{h}) + \inf_\lambda\Big\{\lambda f'_m(0;\bar{h}) + \frac{\lambda^2}{2n}\sum_{s=1}^{n} W''\Big(Y_i\big(F_m(x_s) + \tilde{\lambda}(\lambda)\bar{h}(x_s)\big)\Big)\Big\}\end{aligned}$$

where $\tilde{\lambda}(\lambda)$ lies between 0 and $\bar{\lambda}$. By condition 01,

$$\begin{aligned}&\inf_\lambda\Big\{\lambda f'_m(0;\bar{h}) + \frac{\lambda^2}{2n}\sum_{s=1}^{n} W''\Big(Y_i\big(F_m(x_s) + \tilde{\lambda}(\lambda)\bar{h}(x_s)\big)\Big)\Big\}\\ &\leq \inf_\lambda\Big\{\lambda f'_m(0;\bar{h}) + \frac{\lambda^2\kappa}{4n}\sum_{s=1}^{n} W\Big(Y_i\big(F_m(x_s) + \tilde{\lambda}(\lambda)\bar{h}(x_s)\big)\Big) + \frac{\lambda^2\kappa}{4}\Big\}\\ &\leq \inf_\lambda\{\lambda f'_m(0;\bar{h}) + \frac{\lambda^2\kappa}{2}\}\end{aligned} \tag{5}$$

because $\frac{1}{n}\sum_{s=1}^{n} W(Y_i(F_m(x_s)+\widetilde{\lambda}(\lambda)\bar{h}(x_s)) \le \frac{1}{n}\sum_{s=1}^{n} W(Y_iF_m(x_s)) \le W(0)=1$ since $\bar{\lambda}$ minimizes $f_m(\lambda;\bar{h})$ on $[0,\bar{\lambda}]$, $\widetilde{\lambda}$ is an intermediate point, and $F_0 \equiv 0$. Combining (4) and (5) and the minimizing property of $\underline{h}$,

$$\begin{aligned} f_m(\bar{\lambda};\bar{h}) &\le f_m(0;\bar{h}) - \frac{\left(f_m'(0;\bar{h})\right)^2}{2\kappa} \\ &\le f_m(0;\bar{h}) - \frac{1}{2\kappa(n\sum_{j=1}^{k}\alpha_j)^2} \quad . \end{aligned}$$

The second statement of the theorem follows because the initial value of $n^{-1}\sum_{i=1}^{n} W\left(Y_iF_0(x_i)\right)$ is 1, and the value would fall below 0 after at most $m = 2\kappa(n\sum_{j=1}^{k}\alpha_j)^2$ steps in which at least one observation is not classified correctly. Since the value is necessarily positive, we conclude that all observations would be classified correctly before the mth step.

∎

6. Achieving Rates with Sieve Boosting

We propose a regularization of L_2 boosting which we view as being in the spirit of the original proposal, but, unlike it, can be shown for, suitable $\mathcal{H}$, to achieve minimax rates for estimation of $E(Y|X)$ under quadratic loss for $\mathcal{P}$ for which $E(Y|X)$ is assumed to belong to a compact set of functions such as a ball in Besov space if $X \in \mathbb{R}$ or to appropriate such subsets of spaces of smooth functions in $X \in \mathbb{R}^d$—see, for example, the classes $\mathcal{F}$ of Györfi et al. (2003). In fact, they are adaptive in the sense of Donoho et al (1995) for scales of such spaces. We note that Bühlmann and Yu (2003) have introduced a version of L_2 boosting which achieves minimax rates for Sobolev classes on $\mathbb{R}$ adaptively already. However, their construction is in a different spirit than that of most boosting papers. They start out with $\mathcal{H}$ consisting of one extremely smooth and complex function and show that boosting reduces bias (roughness of the function) while necessarily increasing variance. Early stopping is still necessary and they show it can achieve minimax rates.

It follows, using a result of Yang (1999) that our rule is adaptive minimax for classification loss for some of the classes we have mentioned as well. Unfortunately, as pointed out by Tsybakov (2001), the sets $\{x : |F_B(x)| \le \varepsilon\}$ can behave very badly as $\varepsilon \downarrow 0$, no matter how smooth F_B, the misclassification Bayes rule, is, so that these results are not as indicative as we would like them to be. In a recent paper, Bartlett, Jordan, and McAuliffe (2003) considered minimization of the W empirical risk $n^{-1}\sum_{i=1}^{n} W(Y_iF(X_i))$, for fairly general convex W, over sets of the form $\mathcal{F} = \{F = \sum_{j=1}^{m}\alpha_j h_j,\ h_j \in \mathcal{H},\ \sum_{j=1}^{m}|\alpha_j| \le \alpha_n,$ (for some representation of F)$\}$. They obtained oracle inequalities relating $EW(Y\hat{F}(X))$ for $\hat{F}_j$ the empirical minimizer over $\mathcal{F}_j$ to the empirical W risk minimum. They then proceeded to show using conditions related to Tsybakov's (A1) above how to relate the misclassification regret of $\hat{\mathcal{F}}_j$, given by $\langle P[Y\hat{F}_j(X)<0] - P[YF_B(X)<0]\rangle$ to $\langle E_pW(Y\hat{F}_j) - E_pW(YF_B^*)\rangle$, the W regret where F_B^* is the Bayes rule for W. Using these results (Theorems 3 and 10) they were able to establish oracle inequalities for $\hat{F}_j$ under misclassification loss. Manor, Meir, and Zhang (2004) considered the same problem, but focused their analysis mainly on L_2 boosting. They obtained an oracle inequality similar to that of Bartlett et al. regularizing by permitting step sizes which are only a fraction $\beta < 1$ of the step size declared optimal by Gauss-Southwell. They went further by obtaining near minimax results on suitable sets.

We also limit our results to L_2 boosting, although we believe this limitation is primarily due to the lack of minimax theorems for prediction when other losses than L_2 are considered. We use yet a different regularization method in what follows. We show in Theorem 8 our variant of L_2 boosting achieves minimax rates for estimating $E(Y|X)$ in a wide class of situations. Boosting up to a simple data-determined cutoff in each sieve level of a model, and then cross-validating to choose between sieve levels, we can obtain results equivalent to those in which full optimization using penalties are used, such as Theorem 2.1 of Baraud (2000) and results of Baron, Birgé, Massart (1999). Then, in Theorem 9, we show, using inequalities related to ones of Tsybakov (2001), Zhang (2004) and Bartlett et al. (2003), that the rules we propose are also minimax for 0–1 loss in suitable spaces.

6.1 The Rule

Our regularization requires that $\mathcal{H} \equiv \mathcal{H}^{(\infty)} = \overline{\cup_{m\geq 1}\mathcal{H}^{(m)}}$ where $\mathcal{H}^{(m)}$ are finite sets with certain properties. For instance, if $\mathcal{H}$ consists of the stumps in $[0,1]$, $\mathcal{H} = \{F_y(\cdot) : F_y(x) = \mathrm{sgn}(x-y),\ x,y \in [0,1]\}$ we can take $\mathcal{H}^{(m)} = \{F_y(\cdot) : y$ a dyadic number of order k, $y = \frac{j}{2^k}$, $0 \leq j \leq 2^k\}$. Essentially, we construct a sieve approximating $\mathcal{H}$. Let $\mathcal{F}^{(m)}$ be the linear span of $\mathcal{H}^{(m)}$. Evidently $\mathcal{F} = \overline{\cup_{m\geq 1}\mathcal{F}^{(m)}}$. Let $|\mathcal{H}^{(m)}| \equiv D_m$. Then, $\dim(\mathcal{F}^{(m)}) = D_m$. We now describe our proposed regularization of L_2 boosting.

We use the following notation of Section 4, and begin with a glossary and conditions. Let $(X_1,Y_1),\ldots,(X_n,Y_n),(X,Y)$ i.i.d. with

$$
\begin{aligned}
(X,Y) &\sim P << \mu, \quad P \in \mathcal{P}, \quad \mathbf{X} \equiv (X_1,\ldots,X_n),\ \mathbf{Y} \equiv (Y_1,\ldots,Y_n)\,. \\
Y &\in \{-1,1\} \\
\|f\|_\mu^2 &\equiv \int f^2\, d\mu \\
\|f\|_n^2 &\equiv \frac{1}{n}\sum_{i=1}^n f^2(X_i,Y_i) \\
\|f\|_\infty &= \sup_{x,y} |f(x,y)| \\
F_P(X) &\equiv E_P(Y|X) \\
\hat{F}_m(X) &= \arg\min\{\|t(X)-Y\|_n^2 : t \in \mathcal{F}^{(m)}\} \\
F_m(X) &= \arg\min\{\|t(X)-Y\|_P^2 : t \in \mathcal{F}^{(m)}\} \\
E_{\mathbf{X}} &\equiv \text{Conditional expectation given } X_1,\ldots,X_n
\end{aligned}
$$

Note that we will often suppress $\mathbf{X},\mathbf{Y}$ in $v(\mathbf{X},\mathbf{Y},X,Y)$ and drop subscript to P.

Let $\hat{F}_{m,k}$, the kth iterate in $\mathcal{F}_m$, be defined as follows

$$
\begin{aligned}
\hat{F}_{1,0} &\equiv F_0 \\
\hat{F}_{m+1,0} &= \hat{F}_{m,\hat{k}(m)} \\
\hat{F}_{m,k+1} &= \hat{F}_{m,k} + \hat{\lambda}_{m,k}\hat{h}_{m,km}
\end{aligned}
$$

where

$$
\begin{aligned}
(\hat{\lambda}_{m,k},\hat{h}_{m,k}) &\equiv \arg\min_{\lambda\in\mathbb{R},h\in\mathcal{H}^{(m)}} \{-2\lambda P_n(Y-\hat{F}_{m,k})h + \lambda^2 P_n(h^2)\} \\
\hat{k}(m) &= \text{First } k \text{ such that } \hat{\lambda}_{m,k}^2 \leq \Delta_{m,n},
\end{aligned}
$$

where $\Delta_{m,n}$ are constants. Let

$$\tilde{F}_m = H(\hat{F}_{m,\hat{k}(m)})$$

where

$$H(x) = \begin{cases} x & \text{if } |x| \le 1 \\ \operatorname{sgn}(x) & \text{if } |x| > 1 \end{cases} \tag{6}$$

Note that we have suppressed dependence on n here, indicating it only by the "hats". Let,

$$\hat{m} = \arg\min\{\|Y - \widetilde{F}_m(x)\|_* :\ m \le M_n\}$$

where

$$\|f\|_*^2 = \frac{1}{B}\sum_{i=n+1}^{n+B} f^2(X_i, Y_i), \text{ and we take } B = B_n = \frac{n}{\log n}\ .$$

The rule we propose is: $\hat{\delta} = \operatorname{sgn}(\hat{\hat{F}})$, where

$$\hat{\hat{F}} \equiv H(F_{\hat{m},\hat{k}(\hat{m})})\ . \tag{7}$$

Note: We show at the end of the Appendix (Proof of Lemma 10) that for wavelet $\mathcal{H}$ we take at most $Cn\log n$ steps total in this algorithm.

6.2 Conditions and Results

We use C as a generic constant throughout, possibly changing from line to line but not depending on m, n, or P. Lemma 6.3 and the condition we give are essentially due to Baraud (2001). Let μ be a sigma finite measure on $\mathcal{H}$ and $\|f\|_\mu$ be the $L_2(\mu)$ norm.

R1. If $\mathcal{H}^{(m)} = \{h_{m,1}, \ldots, h_{m,D_m}\}$ and $f_{m,j} \equiv h_{m,j}/\|h_{m,j}\|_\mu$, then $\{f_{m,j}\}, j \ge 1$ is an orthonormal basis of $\mathcal{F}^{(m)}$ in $L_2(\mu)$ such that:

(i) $\|f_{m,j}\|_\infty \le C_\infty D_m^{\frac{1}{2}}$ for all j, where $\|f\|_\infty = \sup_x |f(x)|$.

(ii) There exists an L such that for all m, j, j',
$f_{m,j} f_{m,j'} = 0$ if $|j - j'| \ge L$.

R2. There exists $\varepsilon = \varepsilon(P) > 0$ such that, $\varepsilon \le \frac{dP}{d\mu} \le \varepsilon^{-1}$ for all $P \in \mathcal{P}$.

R3. $\sup_{P\in\mathcal{P}} \|F_P - F_m\|_P^2 \le C D_m^{-\beta}$ for all m, $\beta > 1$.

R4. $M_n \le D_{M_n} \le \frac{n}{(\log n))^p}$ for some $p > 1$.

Condition R1 is needed to conclude that we can bound the behavior of the L_∞ norm on $\mathcal{F}^{(m)}$ by that of the L_2 norm for μ. Condition R2 simply ensures that we can do so for $P \in \mathcal{P}$ as well. The members $f_{m,j}$ of the basis of $\mathcal{F}^{(m)}$ must have compact support. It is well known that if $\mathcal{H}_m$ consists of scaled wavelets (in any dimension) then R1 holds. Clearly, if say μ is Lebesgue measure on an hypercube then to satisfy R2 $\mathcal{P}$ can consist only of densities bounded from above and away from 0. Condition R3 gives the minimum approximation error incurred by using an estimate F based on $\mathcal{F}^{(m)}$, and thus limits our choice of $\mathcal{H}$. Finally, R4 links the oracle error for these sequences of procedures to the number of candidate procedures.

Let

$$r_n(P) = \inf\{E_P\|\hat{F}_m - F_P\|_P^2 : 1 \le m \le M_n\}, \qquad r_n \equiv \sup_{P\in\mathcal{P}} r_n(P).$$

Thus, r_n is the minimax regret for an oracle knowing P but restricted to $\hat{F}_m$. We use the notation $a_n \asymp b_n$ for a shortcut for $a_n = O(b_n)$ and $b_n = O(a_n)$, We have

Theorem 8 *Suppose that $\mathcal{P}$ and $\mathcal{F}$ satisfy* R1–R4 *and that $\mathcal{H}$ is a VC class. If $\Delta_{m,n} = O(D_m/n)$, then,*

$$\sup_{\mathcal{P}} E_P\|\hat{\hat{F}}(X) - F_P(X)\|_P^2 \asymp r_n \,. \tag{8}$$

Thus, $\hat{\hat{F}}$ given by (7) is rate minimax.

Theorem 9 *Suppose the assumptions of Theorem 8 hold and $\mathcal{P}_0 = \mathcal{P} \cap \{P : P\big(\,|F_P(X)| \le t\big) \le ct^\alpha\}$, $\alpha \ge 0$. Let $\Delta_n(F,P)$ be the Bayes classification regret for P,*

$$\Delta_n(F,P) \equiv P\big(YF(X) < 0\big) - P\big(YF_P(X) < 0\big)\,. \tag{9}$$

Then,

$$\sup_{\mathcal{P}_0} \Delta_n(\hat{\hat{F}},P) \asymp r_n^{\frac{\alpha+1}{\alpha+2}}\,. \tag{10}$$

The condition $P[|F_P(x)| \le t] \le ct^\alpha$, some $\alpha \ge 0$, t sufficiently small appears in Proposition 1 of Tsybakov (2001) as sufficient for his condition (A1) which is studied by both Bartlett et al. (2003) and Mammen and Tsybakov (1999).

The proof of Theorem 9 uses 2 lemmas of interest which we now state. Their proofs are in the Appendix.

We study the algorithm on $\mathcal{F}_m$. For any positive definite matrix Σ define the condition number $\gamma(\Sigma) \equiv \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$, where $\lambda_{\max}$, $\lambda_{\min}$ are the largest and smallest eigenvalues of Σ. Let $G_m(P) = \|E_P f_{m,i}(X) f_{m,j}(X)\|$ be the $D_m \times D_m$ Gram matrix of the basis $\{f_{m,1},\dots,f_{m,D_m}\}$.

Lemma 10 *Under* R1 *and* R2*,*

a) $\gamma(G_m(P)) \le \varepsilon^{-2}$, where ε is as in R2.

b) Let $G_m(P_n)$ be the empirical Gram matrix $\hat{\gamma}_m \equiv \gamma(G_m(P_n))$. Then, if in addition to R1 *and* R2*, $\mathcal{H}$ is a VC class, $P[\gamma(\hat{G}_m) \ge C_1] \le C_2 \exp\{-C_3 n/L^2 D_m\}$ for all $m \le M_n$ for such that $D_m \le n/(\log n)^p$ for $p > 1$.*

c) If $\mathcal{H}$ is a VC class, $P[\|\hat{F}_{m,\hat{k}(m)} - \hat{F}_m\|_k \le C\frac{D_m}{n}] = 1 - O(\frac{1}{n})$ The C and 0 terms are determed solely by the constants appearing in the R conditions.

Lemma 11 *Suppose* R1, R2*, and* R4 *hold. Then,*

$$E_P(\widetilde{F}_m - F_P)^2 \le C\{E_P(F_m - F_P)^2 + \frac{D_m}{n} + E_P(\tilde{F}_m - \hat{F}_m)^2\}.$$

This "oracle inequality" is key for what follows.

Proof of Theorem 9

$$P\big(YF(x)<0\big)=\frac{1}{2}E_P\Big(1(F(X)>0)\big(1-F_P(X)\big)\Big)+\frac{1}{2}E_P\Big(1(F(X)<0)\big(1+F_P(X)\big)\Big).$$

Hence for all $\varepsilon>0$,

$$\begin{aligned}\Delta_n(F,P)&=E_P\Big(1\big(F(X)<0,F_P(X)>0\big)F_P(X)-1\big(F(X)>0,F_P(X)<0\big)F_P(X)\Big)\\&=E_P\Big(|F_P(X)|1(F_P(X)F(X)<0)\Big)\\&\le E_P\Big(\big|F(X)-F_P(X)\big|1(F_PF(X)<0,|F_P(X)|>\varepsilon)\Big)+\varepsilon P\big(|F_P(X)|\le\varepsilon\big)\\&\le\frac{1}{\varepsilon}E_P\big(F(X)-F_P(X)\big)^2+c\varepsilon^{\alpha+1}\end{aligned}$$

by assumption. The theorem follows. ■

6.3 Discussion

1) If $X\in\mathbb{R}$ and $\mathcal{H}^{(m)}$ consists of stumps with the discontinuity at a dyadic rational $j/2^m$, then $\mathcal{F}^{(m)}$ is the linear space of Haar wavelets of order m. This is also true if $\mathcal{H}_m$ is the space of differences of two such dyadic stumps. More generally, if $\mathcal{H}$ consists of suitably scaled wavelets, so that $|h|\le 1$, based on the dyadic rationals of order m, them $\mathcal{F}^{(m)}$ is the linear space spanned by the first 2^m elements of the wavelet series. A slight extension of results of Baraud (2001) yields that if we run the algorithm to the limit $k=\infty$ for each m rather than stopping as we indicate, the resulting $\hat{F}_m$ obey the oracle inequality of Lemma 11 with $\Delta_{m,n}=0$.

   Suppose that $X\in\mathbb{R}$ and F_∞ ranges over a ball in an approximation space such as Sobolev or, more generally, Besov. Then, if $\mathcal{F}^{(m)}$ has the appropriate approximation properties, e.g., wavelets as smooth as the functions in the specified space, it follows from Baraud (2001) that we can use penalties not dependent on the data to pick $\hat{F}_{\hat{m}}$ such that,

$$\begin{aligned}\max_{\hat{F}}E_P\Big(\hat{F}_{\hat{m}}(X)-E_P(Y|X)\Big)^2&\asymp\min_{\hat{F}}\max\Big\{E_P\big(\hat{F}(X)-E_P(Y|X)\big)^2:E_P(Y|X)\in\mathcal{F}\Big\}\\&\asymp n^{-1+\varepsilon}\Omega(n)\end{aligned}$$

   where $\Omega(n)$ is slowly varying and $0<\varepsilon<1$. Here $\hat{F}$ ranges over all estimators based only on the data and not on P. The same type of result has been established for more specialized models with $X\in\mathbb{R}^d$ by Baron, Birgé, Massart (1999), and others, see Györfi et al. (2003).

   The resulting minimax risk,

$$\min_{\hat{F}}\max\{E_P\big(\hat{F}(X)-E_P(Y|X)\big)^2:\ E_P(Y|X)\in\mathcal{F}\}$$

   is always of order $n^{-1+\varepsilon}\Omega(n)$ where $\Omega(n)$ is typically constant and $0<\varepsilon<1$.

   What we show in Theorem 8 is that if, rather than optimizing all the way for each m, we stop in a natural fashion and cross validate as we have indicated, then we can achieve the optimal order as well.

2) "Stumps" unfortunately do not satisfy condition R1 with μ Lebesgue measure. Their Gram matrices are too close to being singular. But differences of stumps work.

3) It follows from the results of Yang (1999) that the rate of Theorem 9 for $\alpha = 0$, that is, if $\mathcal{P}_0 = \mathcal{P}$, is best possible for Sobolev balls and the other spaces we have mentioned.

Tsybakov implicitly defines a class of F_P for which he is able to specify classification minimax rates. Specifically let $X \in [0,1]^d$ and let $b(x_1,\ldots,x_{d-1})$ be a function having continuous partial derivatives up to order ℓ. Let $p_{b,x}(\cdot)$ be the Taylor polynomial or order ℓ obtained from expanding b at x. Then, he defines $\Sigma(l,L)$ to be the class of all such b for which, $|b(y)-p_{b,x}(y)| \leq L|y-x|^\ell$ for all $x,y \in [0,1]^{d-1}$. Evidently if b has bounded partial derivatives of order $\ell+1$, $b \in \Sigma(\ell,L)$, for some L. Now let

$$\begin{aligned}\mathcal{P}_\ell = \quad & \{P : F_P(x) = x_d - b(x_1,\ldots,x_{d-1}),\\ & P[|F_P(x)| \leq t] \leq Ct, \text{ for all } 0 \leq t \leq 1, b \in \Sigma(\ell,L)\}\end{aligned}$$

Tsybakov following Mammen and Tsybakov (1999) shows that the classification minimax regret for $\mathcal{P}$ (Theorem 2 of Tsybakov (2001) for $K = 2$) is $\frac{2\ell}{3\ell+(d-1)}$. On the other hand, if we assume that $Y = F_P(x) + \varepsilon$ where ε is independent of X, bounded and $E(\varepsilon) = 0$, then the L_2 minimax regret rate is $2\ell/(2\ell+(d-1))$ – see Birgé and Massart (1999) Sections 4.1.1 and Theorem 9. Our theorem 9 now yields a classification minimax regret rate of

$$\frac{2}{3}\cdot\frac{2\ell}{2\ell+(d-1)} = \frac{2\ell}{3\ell+\frac{3}{2}(d-1)}$$

which is slightly worse than what can be achieved using Tsybakov's not as readily computable procedures. However, note that as $\ell \to \infty$ so that F_P and the boundary become arbitrarily smooth, L_2 boosting approaches the best possible rate for $\mathcal{P}_\ell$ of $\frac{2}{3}$. Similar remarks can be made about $0 < \alpha \leq 1$.

7. Conclusions

In this paper we presented different mathematical aspects of boosting. We consider the observations as an i.i.d. sample from a population (i.e., a distribution). The boosting algorithm is a Gauss-Southwell minimization of a classification loss function (which typically dominates the 0-1 misclassification loss). We show that the output of the boosting algorithm follows the theoretical path as if it were applied to the true distribution of the population. Since early stopping is possible as argued, the algorithm, supplied with an appropriate stopping rule, is consistent.

However, there are no simple rate results other than those of Bühlmann and Yu (2003), which we discuss, for the convergence of the boosting classifier to the Bayes classifier. We showed that rate results can be obtained when the boosting algorithm is modified to a cautious version, in which at each step the boosting is done only over a small set of permitted directions.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (NSF grant DMS-0104075), and the Israel Science Foundation (ISF grant 793/03).

Appendix A. Proof of Theorem 1:

Let $w_0 = \inf_{f\in\mathcal{F}_\infty} w(f)$. Let $f_k^* = \sum_m \alpha_{km} h_{km}$, $h_{k,m} \in \mathcal{H}$, $\sum_m |\alpha_{km}| < \infty$, $k = 0, 1, 2, \ldots$ be any member of $\mathcal{F}_\infty$ such that (i) $f_0^* = f_0$; (ii) $w(f_k^*) \searrow w_0$ is strictly decreasing sequence; (iii) The following condition is satisfied:

$$w(f_k^*) \geq \alpha w_0 + (1-\alpha) w(f_{k-1}^*) + (1-\alpha)(\nu_{k-1} - \nu_k), \tag{11}$$

where $\nu_k \searrow 0$ is a strictly decreasing real sequence. The construction of the sequence $\{f_k^*\}$ is possible since, by assumption, $\mathcal{F}_\infty$ is dense in the image of $w(\cdot)$. That is, we can start with the sequence $\{w(f_k^*)\}$, and then look for suitable $\{f_k^*\}$. Here is a possible construction. Let c and η be suitable small number. Let $\gamma = (1-\alpha)(1+2\eta)/(1-\eta)$, $\nu_k = c\eta\gamma^k/(1-\gamma)$. Select now f_k^* such $w_0 + c(1-\eta)\gamma^k \leq w(f_k^*) \leq w_0 + c(1+\eta)\gamma^k$. ($\eta$ should be small enough such that $\gamma < 1$ and c should selected such that $w(f_1^*) < w(f_0)$.) Our argument rests on the following,

Lemma 12 *There is a sequence $m_k \to \infty$ such that $w(f_m) \leq w(f_k^*) + \nu_k$ for $m \geq m_k$, $k = 1, 2, \ldots$, and $m_k \leq \zeta_k(m_{k-1}) < \infty$, where $\zeta_k(\cdot)$ is a monotone non-decreasing functions which depends only on the sequences $\{\nu_k\}$ and $\{f_k^*\}$.*

Proof of Lemma 12:

We will use the following notation. For $f \in \mathcal{F}_\infty$ let $\|f\|_* = \inf\{\sum|\gamma_i|,\ f = \sum \gamma_i h_i,\ h_i \in \mathcal{H}\}$.

Recall that by definition $w(f_0) = w(f_0^*)$. Our argument proceeds as follows, We will inductively define m_k satisfying the conclusion of the lemma, and make, if $\varepsilon_{k,m} \equiv w(f_m) - w(f_k^*)$,

$$\varepsilon_{k,m} \leq c_{k,m} \equiv \max\left\{\nu_k,\ \frac{\sqrt{512}B}{\alpha^2\beta_k}\frac{w(f_{k-1}^*) - w_0}{\left(\log\left(1 + \frac{8(w(f_{k-1}^*)-w_0)}{\alpha\beta_k(\tau_k+\rho_k m_{k-1})}(m - m_{k-1}+1)\right)\right)^{1/2}}\right\}, \tag{12}$$

where

$$\begin{aligned} \beta_k &= \inf\{w''(f;h) : w_0 + \nu_k \leq w(f) \leq w(f_0),\ h \in \mathcal{H}\} \\ B &= \sup\{w''(f;h) : w(f) \leq w(f_o),\ h \in \mathcal{H}\} < \infty. \end{aligned} \tag{13}$$

and

$$\begin{aligned} \tau_k &= 2\|f_0 - f_k^*\|_*^2 \\ \rho_k &= \frac{16}{\alpha\beta_k}\big(w(f_0) - w_0\big). \end{aligned} \tag{14}$$

Having defined m_k we establish (12) as part of our induction hypothesis for $m_{k-1} < m \leq m_k$. We begin by choosing $m = m_1 = 1$ so that (12) holds for $m = M - 1 = 1$. We do do this by choosing $\nu_0 > 0$, sufficiently small. Having established the induction for $m \leq m_{k-1}$ we define m_k as follows. Write now the RHS of (12) as $g(m_{k-1})$, where

$$g(\nu) \equiv \max\left\{\nu_k,\ \frac{\sqrt{512}B}{\alpha^2\beta_k}\frac{w(f_{k-1}^*) - w_0}{\left(\log\left(1 + \frac{8(w(f_{k-1}^*)-w_0)}{\alpha\beta_k(\tau_k+\rho_k \nu}(m - \nu+1)\right)\right)^{1/2}}\right\},$$

We can now pick $\zeta_k(\nu) \equiv \max\{\nu+1, \min\{m: g(\nu) \le \nu_k\}\}$, and define $m_k = \zeta_k(\nu_{k-1})$.

Note that $\{\beta_k\}$, $\{\tau_k\}$, $\{\rho_k\}$, and B depend only the sequences $\{f_k^*\}$ and $\{\nu_k\}$. We now proceed to establish (12). for $m_{k-1} < m \le m_k$. Note first that since $\varepsilon_{k,m}$ as a function of m is non-increasing, (12) holds trivially for $m' > m$ if $\varepsilon_{k,m} \le 0$. By induction (12) holds for $m \le m_{k-1}$, and my hold for some $m > mk-1$. Recall that the definition of the algorithm relates the actual gain at the mth to the maximal gain achieved in this step given the previous steps, see its definition (1). Suppose

$$\inf_\lambda w(f_m + \lambda h_m) \le w_0 + \nu_k. \tag{15}$$

Then

$$\begin{aligned}
w(f_{m+1}) &\le \alpha \inf_\lambda w(f_m + \lambda h_m) + (1-\alpha) w(f_m), \quad \text{by (1)} \\
&\le \alpha(w_0 + \nu_k) + (1-\alpha) w(f_m), \quad \text{by (15)} \\
&\le \alpha(w_0 + \nu_k) + (1-\alpha)\big(w(f_{k-1}^*) + \nu_{k-1}\big), \quad \text{by the outer induction, since } m \ge m_{k-1} \\
&\le \alpha(w_0 + \nu_k) + \big(w(f_k^*) - \alpha w_0 + (1-\alpha)\nu_k\big), \quad \text{by (11)} \\
&= w(f_k^*) + \nu_k,
\end{aligned}$$

so that $\varepsilon_{k,m+1} \le \nu_k$. Therefore, m_k' is not larger than $m+1$, that is $\varepsilon_{k,m'} \le \nu_k$ for $m' > m$ then (12) holds trivially for $m' > m$, and hence, by the second induction assumption for all m. We have established (12) save for m such that,

$$\inf_\lambda w(f_m + \lambda h_m) > w_0 + \nu_k \text{ and } \varepsilon_{k,m} \ge 0. \tag{16}$$

We now deal with this case.

Note first that by convexity,

$$|w'(f_m; f_m - f_k^*)| \ge w(f_m) - w(f_k^*) \equiv \varepsilon_{k,m}. \tag{17}$$

We obtain from (17) and the linearity of the derivative that, if $f_m - f_k^* = \sum \gamma_i \tilde{h}_i \in \mathcal{F}_\infty$,

$$\varepsilon_{k,m} \le \left| \sum -\gamma_i w'(f_m; \tilde{h}_i) \right| \le \sup_{h \in \mathcal{H}} |w'(f_m; h)| \sum |\gamma_i| \, .$$

Hence

$$\sup_{h \in \mathcal{H}} |w'(f_m; h)| \ge \frac{\varepsilon_{k,m}}{\|f_m - f_k^*\|_*}. \tag{18}$$

Now, if $f_{m+1} = f_m + \lambda_m h_m$ then,

$$w(f_m + \lambda_m h_m) = w(f_m) + \lambda_m w'(f_m; h_m) + \frac{1}{2}\lambda_m^2 w''(\tilde{f}_m; h_m), \quad \lambda \in [0, \lambda_m]. \tag{19}$$

where $\tilde{f}_m = f_m + \tilde{\lambda}_m h_m$ and $0 \le \tilde{\lambda}_m \le \lambda_m$. By convexity, for $0 \le \lambda \le \lambda_m$,

$$w(f_m + \lambda h_m) = w(f_m(1 - \frac{\lambda}{\lambda_m}) + \frac{\lambda}{\lambda_m} f_{m+1}) \le \max\{w(f_m), w(f_{m+1})\} = w(f_m) \le w(f_1).$$

We obtain from Assumption GS1 that $w''(\tilde{f}_m;h) \in (\beta_k, B)$ given in (13). But then we conclude from (19) that,

$$\begin{aligned} w(f_m+\lambda_m h_m) &\geq w(f_m) + \inf_{\lambda\in\mathbb{R}}\big(\lambda w'(f_m;h_m) + \frac{1}{2}\lambda^2\beta_k\big) \\ &= w(f_m) - \frac{|w'(f_m;h_m)|^2}{2\beta_k} \ . \end{aligned} \tag{20}$$

Note that $w(f_m+\lambda h) = w(f_m)+\lambda w'(f_m,h)+\lambda^2 w''(f_m+\lambda' h,h)/2$ for some $\lambda' \in [0,\lambda]$, and if $w(f_m+\lambda h)$ is close to $\inf_{\lambda,h} w(f_m+\lambda, h)$ then by convexity, $w(f_m+\lambda' h) \leq w(f_m) \leq w(f_0)$. We obtain from the upper bound on w'' we obtain:

$$\begin{aligned} w(f_m+\lambda_m h_m) &\leq \alpha \inf_{\lambda\in\mathbb{R},\, h\in\mathcal{H}} w(f_m+\lambda h) + (1-\alpha)w(f_m), \qquad \text{by definition,} \\ &\leq \alpha \inf_{\lambda\in\mathbb{R},\, h\in\mathcal{H}} (w(f_m)+\lambda w'(f_m;h)+\frac{1}{2}\lambda^2 B) + (1-\alpha)w(f_m) \\ &= w(f_m) - \frac{\alpha \sup_{h\in\mathcal{H}} |w'(f_m;h)|^2}{2B}, \end{aligned} \tag{21}$$

by minimizing over λ. Hence combining (20) and (21) we obtain,

$$|w'(f_m;h_m)| \geq \alpha \sup_{h\in\mathcal{H}} |w'(f_m;h)| \sqrt{\frac{\beta_k}{B}} \tag{22}$$

By (21) for the LHS and convexity for the RHS:

$$\frac{\alpha \sup_{h\in\mathcal{H}} |w'(f_m;h)|^2}{2B} \leq w(f_m) - w(f_{m+1}) \leq -\lambda_m w'(f_m;h_m)$$

Hence

$$|\lambda_m| \geq \frac{\alpha \sup_{h\in\mathcal{H}} |w'(f_m;h)|}{2B}.$$

Applying (18) we obtain:

$$|\lambda_m| \geq \frac{\alpha}{2B}\frac{\varepsilon_{k,m}}{l_{k,m}}, \tag{23}$$

where $l_{k.m} \equiv \|f_m - f_k^*\|_*$.

Let λ_m^0 be the minimal point of $w(f_m+\lambda h_m)$. Taylor expansion around that point and using the lower bound on the curvature:

$$w(f_m+\lambda h_m) \geq w(f_m+\lambda_m^0 h_m) + \frac{1}{2}\beta_k(\lambda-\lambda_m^0)^2 \tag{24}$$

Hence

$$\begin{aligned} {\lambda_m^0}^2 &\leq \frac{2}{\beta_k}\big(w(f_m) - w(f_m+\lambda_m^0 h_m)\big) \\ &\leq \frac{2}{\alpha\beta_k}\big(w(f_m) - w(f_{m+1})\big), \end{aligned} \tag{25}$$

where the RHS follows (1). Similarly

$$
\begin{aligned}
(\lambda_m - \lambda_m^0)^2 &\le \frac{2}{\beta_k}\big(w(f_{m+1}) - w(f_m + \lambda_m^0 h_m)\big) \\
&\le \frac{2(1-\alpha)}{\alpha\beta_k}\big(w(f_m) - w(f_{m+1})\big)
\end{aligned}
\tag{26}
$$

Combining (25) and (26):

$$
\lambda_m^2 \le \frac{8}{\alpha\beta_k}\big(w(f_m) - w(f_{m+1})\big). \tag{27}
$$

Since $\varepsilon_{k,m} \ge 0$ by assumption (16), we conclude from (27) that,

$$
\sum_{i=m_{k-1}}^{m} \lambda_i^2 \le \frac{8}{\alpha\beta_k}\big(w(f_{k-1}^*) - w_0\big). \tag{28}
$$

However, by definition,

$$
\begin{aligned}
l_{k,m+1} &\le l_{k,m} + |\lambda_m| \\
&\le l_k + \sum_{i=m_{k-1}}^{m} |\lambda_i| \\
&\le l_k + (m+1-m_{k-1})^{1/2}\Big(\sum_{i=m_{k-1}}^{m} \lambda_i^2\Big)^{1/2}
\end{aligned}
\tag{29}
$$

by Cauchy-Schwarz, where, similarly,

$$
\begin{aligned}
l_k = l_{k,m_{k-1}} &= \|f_{m_{k-1}} - f_k^*\|_* \\
&\le \|f_0 - f_k^*\|_* + \|f_{m_{k-1}} - f_0\|_* \\
&\le \|f_0 - f_k^*\|_* + \sum_{m=0}^{m_{k-1}-1} |\lambda_m| \\
&\le \|f_0 - f_k^*\|_* + m_{k-1}^{1/2}\sqrt{\sum_{m=0}^{m_{k-1}-1} \lambda_m^2} \\
&\le \|f_0 - f_k^*\|_* + \sqrt{\frac{8m_{k-1}}{\alpha\beta_k}}\sqrt{w(f_0) - w(f_{m_{k-1}})}, \quad \text{by (27)} \\
&\le \|f_0 - f_k^*\|_* + \sqrt{\frac{8m_{k-1}}{\alpha\beta_k}}\sqrt{w(f_0) - w_0} \\
&\le \sqrt{\tau_k + \rho_k m_{k-1}}, \quad \text{as defined in (14).}
\end{aligned}
\tag{30}
$$

Together, (23), (28), and (29) yield:

$$\begin{aligned}\frac{8}{\alpha\beta_k}(w(f^*_{k-1})-w_0) &\geq \sum_{i=m_{k-1}}^{m} \lambda_i^2 \\ &\geq \frac{\alpha^2}{4B^2}\sum_{i=m_{k-1}}^{m}\frac{\varepsilon_{k,i}^2}{l_{k,i}^2} \\ &\geq \frac{\alpha^2}{4B^2}\sum_{i=m_{k-1}}^{m}\frac{\varepsilon_{k,i}^2}{(l_k+(8(w(f^*_{k-1})-w_0)/\alpha\beta_k)^{1/2}(i-m_{k-1})^{1/2})^2}\end{aligned} \tag{31}$$

Further, since $\varepsilon_{k,m}$ are decreasing by construction and positive by assumption (16), we can simplify the sum on the RHS of (31):

$$\begin{aligned}\sum_{i=m_{k-1}}^{m}&\frac{\varepsilon_{k,i}^2}{(l_k+(8(w(f^*_{k-1})-w_0)/\alpha\beta_k)^{1/2}(i-m_{k-1})^{1/2})^2} \\ &\geq \frac{\varepsilon_{k,m}^2}{2}\sum_{i=0}^{m-m_{k-1}}\frac{1}{l_k^2+8i(w(f^*_{k-1})-w_0)/\alpha\beta_k}.\end{aligned} \tag{32}$$

Using the inequality,

$$\sum_{i=0}^{m-m_{k-1}}\frac{1}{a+bi} \geq \int_0^{m-m_{k-1}+1}\frac{1}{a+bt}\,dt = \frac{1}{b}\log\Big(1+\frac{b}{a}(m-m_{k-1}+1)\Big)$$

on the RHS of (32), we obtain from (31) and (32) that (12) holds, for the case (16). This establishes (16) for all k and m.

■

Proof of Theorem 1: Since the lemma established the existence of monotone ζ_k's, it followed from the definition of these function that $w(f_m) \leq w(f^*_{k(m)})$ where $k(m) = \sup\{k : \zeta^{(k)}(f^*_0) \leq m\}$ and $\zeta^{(k)} = \zeta_k \circ \cdots \circ \zeta_1$ is the kth iterate of the ζs. Since $\zeta^{(k)}(f^*_0) < \infty$ for all k, we have established the uniform rate of convergence and can define the sequence $\{c_m\}$, where $c_m = w(f^*_{k(m)}) - w_0$.

We now prove the uniform step improvement claim of the theorem and identify a suitable function $\xi(\cdot)$. From (26) and (23) if $\varepsilon_{k,m} \geq 0$

$$w(f_m) - w(f_{m+1}) \geq \frac{\alpha\beta_k}{2}\lambda_m^2 \geq \frac{\alpha\beta_k}{2}\left(\frac{\alpha}{2B}\frac{\varepsilon_{k,m}}{l_{k,m}}\right)^2, \tag{33}$$

Bound $l_{k,m}$ similarly to (30) by

$$l_{k,m} \leq l_{k,1} + m^{1/2}\Big(\sum_{i=1}^{m}\lambda_i^2\Big)^2 \leq l_{k,1} + \sqrt{\frac{8m}{\alpha\beta_k}(w(f_0)-w_0)}. \tag{34}$$

Let $m^*(v) = \inf\{m' : c_{m'} \leq v - w_0\}$, which is well defined since $c_m \to 0$. Thus, any realization of the algorithm will cross the v line on or before step number $m^*(v)$. In particular, $m \leq m^*\big(w(f_m)\big)$ for any m and any realization of the algorithm. We obtain therefore by plugging-in (34) in (33), using the m^* as a bound on m and the identity $(a+b)^2 \leq 2a^2+2b^2$ that:

$$w(f_m)-w(f_{m+1}) \geq \frac{\alpha^3\beta_k}{16B^2}\frac{w(f_m)-w(f_k^*)}{l_{k,1}^2+8m^*\big(w(f_m)\big)\big(w(f_0)-w_o\big)/\alpha\beta_k},$$

as long as $\varepsilon_{k,m} \geq 0$. Taking the maximum of the RHS over the permitted range, yields a candidate for the ξ function:

$$\xi(w) \equiv \sup_{k:\, w(f_k^*)\leq w}\left\{\frac{\alpha^3\beta_k}{16B^2}\frac{w-w(f_k^*)}{l_{k,1}^2+m^*(w)\big(w(f_0)-w_o\big)/\alpha\beta_k}\right\}.$$

This proves the theorem under GS1. Under GS2, the only inequality which we need to replace is (20) since now $\beta_k = 0$ is possible. However the definition of Algorithm 2 ensures that we have a coefficient of at least γ on λ^2 in (20). The theorem is proved.

■

Appendix B. Proof of Lemmas 10 and 11 and Theorem 8

Proof of Lemma 10 Since by (R2)

$$\begin{aligned}
\lambda_{\max}(G_m(P)) &= \sup_{\|x\|=1} x'G_m(P)x \\
&= \sup_{\|x\|=1} \sum\sum x_i x_j \int f_{m,i}f_{m,d}\,dP \\
&= \sup_{\|x\|=1} \int (\sum x_i f_{m,i})^2 dP \qquad (35)\\
&\leq \varepsilon^{-1} \sup_{\|x\|=1} \int (\sum x_i f_{m,i})^2 d\mu = \varepsilon^{-1} \\
\lambda_{\max}(G_m(P)) &\geq \varepsilon, \quad \text{similarly.}
\end{aligned}$$

Part a) follows.

For any symmetric matrix M define its operator norm $\|\cdot\|_T$ by $\lambda_{\max}(M)$. For simplicity let $G_m = G_m(P)$ and $\hat{G}_m = G_m(P_n)$. Recall that for any symmetric matrices A and and M:

$$\begin{aligned}
|\lambda_{\max}(A)-\lambda_{\max}(M)| &\leq \|A-M\|_T \\
|\lambda_{\min}(A)-\lambda_{\min}(M)| &\leq \|A-M\|_T.
\end{aligned}$$

Now,

$$\begin{aligned}
&P\left[\left|\frac{\lambda_{\max}(\hat{G}_m)}{\lambda_{\min}(\hat{G}_m)}-\frac{\lambda_{\max}(G_m)}{\lambda_{\min}(G_m)}\right| \geq t\right) \\
&\quad \leq P\Big(\|\hat{G}_m-G_m\|_T > \frac{\varepsilon}{2}\Big)+P\Big(\|\hat{G}_m-G_m\|_T \geq t/(\frac{1}{\varepsilon}+\frac{2}{\varepsilon^3})\Big) \qquad (36)
\end{aligned}$$

Recall that for a banded matrix M of with band of width $2L$,

$$\begin{aligned}
\|M\|_T^2 &= \sup_{\|x\|=1} \|Mx\|^2 \\
&= \sup_{\|x\|=1} \sum_a \big(\sum_b M_{ab} x_b\big)^2 \\
&\le \sup_{\|x\|=1} \sum_a \sum_{|b-a|<L} x_b^2 M_\infty^2 \\
&\le 2LM_\infty^2 \sup_{\|x\|=1} \sum_a x_a^2 = 2LM_\infty^2,
\end{aligned}$$

where $\|M\|_\infty \equiv \max_{a,b} |M_{ab}|$. Since both $\hat{G}_m$ and $G_m(P)$ are banded of width d, say,

$$\|\hat{G}_m - G_m\|_T \le 2L \max\Big\{ \Big| \frac{1}{n} \sum_{i=1}^n \big(f_{m,a} f_{m,b})(X_i) - E_P f_{m,a} f_{m,b}(X_i)\big) \Big| \,:\, |a-b| < L \Big\} . \tag{37}$$

If $\mathcal{H}$ is a VC class, we can conclude from (35)–(37) that,

$$P[\gamma(\hat{G}_m) \ge C_1] \le C_2 \exp\{-C_3 n / L^2 D_m\} \tag{38}$$

since by R1 (i), $\|f_m\|_\infty \le C_\infty D_m^{\frac{1}{2}}$. The constants ε, C_1, C_2 and C_3 depend on the constants of the R conditions only. This is a consequence of Theorem 2.14.16 p. 246 of van der Vaart and Wellner (1996). This complete the proof of part b).

By a standard result for the Gauss-Southwell method, Luenberger (1984), page 229:

$$\|\hat{F}_{m,k+1} - \hat{F}_m\|_n^2 \le \Big(1 - \frac{1}{\hat{\gamma}_m D_m}\Big) \|\hat{F}_{m,k} - \hat{F}_m\|_n^2 \tag{39}$$

Hence

$$\|\hat{F}_{m,k} - \hat{F}_m\|_n^2 - \|\hat{F}_{m,k+1} - \hat{F}_m\|_n^2 \ge \frac{1}{\hat{\gamma}_m D_m} \|\hat{F}_{m,k} - \hat{F}_m\|_n^2$$

Thus, if

$$\frac{1}{n} \ge \|\hat{F}_{m,k} - \hat{F}_m\|_n^2 - \|\hat{F}_{m,k+1} - \hat{F}_m\|_n^2$$

we obtain

$$\|\hat{F}_{m,k} - \hat{F}_m\|_n^2 \le D_m \hat{\gamma}_m / n. \tag{40}$$

¿From (40) part (c) follows. ■

Note: Since

$$\|\hat{F}_{m,k-1} - \hat{F}_m\|_n^2 - \|\hat{F}_{m,k} - \hat{F}_m\|_n^2 \;\ge\; \frac{C}{n}$$

(39) implies that

$$\Big(1 - \frac{1}{\hat{\gamma}_m D_m}\Big)^{\hat{k}(m)} \;\ge\; \frac{1}{n}.$$

Therefore:

$$\hat{k}(m) \le \log n \hat{\gamma}_m D_m \,.$$

If, for instance, as with wavelets $D_m = 2^m, m \le \log_2 n$ we take at most $Cn \log n$ steps total.

Lemma 13 *:*
If $E_{\mathbf{x}}$ denotes conditional expectation give n $X_1,\dots,X_n$, under R1 *and $F \equiv F_p$,*

$$E_{\mathbf{x}}\|\hat{F}_m - F_m\|_n^2 \le C(\frac{D_m}{n} + \|F_m - F\|_P^2) \tag{41}$$

This is a standard type of result – see Barron, Birgé, Massart (1999). We include the proof for completeness. Note that,

$$\|\hat{F}_m(X) - Y\|_n^2 = \frac{1}{n}\mathbf{Y}^T(I-P)\mathbf{Y}$$

where $\mathbf{Y} \equiv (Y_1,\dots,Y_n)^T$ and P is the projection matrix of dimension D_m onto the L space spanned by $\big(h_j(X_1),\dots,h_j(X_n)\big)$, $1 \le j \le D_m$. Then, $(I-P)v = 0$ for all $v \in L$. Hence,

$$E_{\mathbf{X}}\|\hat{F}_m(X) - Y\|_n^2 = \frac{1}{n}E_{\mathbf{X}}\big(\mathbf{Y} - \mathbf{F}_m(\mathbf{X})\big)^T(I-P)\big(\mathbf{Y} - \mathbf{F}_m(\mathbf{X})\big)$$

where $\mathbf{F}_m(\mathbf{X}) = \big(F_m(X_1),\dots,F_m(X_n)\big)^T$ is the projection of $(F(X_1),\dots,F(X_n))^T$ onto L. Note also that,

$$\|\hat{F}_m - F_m\|_n^2 = \|\mathbf{Y} - \mathbf{F}_m(\mathbf{X})\|_n^2 - \|\mathbf{Y} - \hat{\mathbf{F}}_m(\mathbf{X})\|_n^2$$

where $\hat{\mathbf{F}}(X) = (\hat{F}_m(X_1),\dots,\hat{F}_m(X_n))^T$. Hence,

$$\begin{aligned}
E_{\mathbf{X}}\|\hat{F}_m - F_m\|_n^2 &= \tfrac{1}{n}E_{\mathbf{X}}(\mathbf{Y} - \mathbf{F}_m(X))^T P(\mathbf{Y} - \mathbf{F}_m(\mathbf{X})) \\
&= \tfrac{1}{n}E_{\mathbf{X}}(\mathbf{Y} - \mathbf{F}(\mathbf{X}))^T P(\mathbf{Y} - \mathbf{F}(\mathbf{X})) + \tfrac{2}{n}E_{\mathbf{X}}(\mathbf{F}_m - \mathbf{F})^T P(\mathbf{Y} - \mathbf{F}_m(\mathbf{X})) \\
&= \tfrac{1}{n}E_{\mathbf{X}}\,\mathrm{trace}[P(\mathbf{Y} - \mathbf{F}(\mathbf{X}))(\mathbf{Y} - \mathbf{F}(\mathbf{X}))] \\
&\quad + \tfrac{2}{n}E_{\mathbf{X}}(\mathbf{F}_m - \mathbf{F})^T P(\mathbf{F}_m - \mathbf{F})(\mathbf{X})
\end{aligned}$$

But

$$E_{\mathbf{X}}\,\mathrm{trace}[P(\mathbf{Y} - \mathbf{F}(\mathbf{X})))(\mathbf{Y} - \mathbf{F}(\mathbf{X}))^T] = \frac{1}{n}\sum_{i=1}^n Var(Y_i|X_i)p_{ii}(X) \le \max_i Var(Y_i|X_i)\frac{D_m}{n}$$

since

$$\sum_{i=1}^n p_{ii}(X) = \mathrm{trace}\, P = D_m$$

Also, since P is a projection matrix

$$(\mathbf{F}_m - \mathbf{F})^T P(\mathbf{F}_m - \mathbf{F})(\mathbf{X}) \le \|F_m - F\|_n^2$$

and (41) *follows.*

Proof of Lemma 11:
Take $\Delta_{m,n} = 0$. Let $\tilde{\rho}_m = \sup\left\{\frac{\|t(X)\|_P}{\|t(X)\|_n} \ : \ t \in \mathcal{F}_m\right\}$. By Proposition 5.2 of Baraud (2001), if $\rho_0 > h_0^{-1}$,

$$P[\tilde{\rho}_m > \rho_0] \le D_m^2 \exp\{-\frac{(h_0 - \rho_0^{-1})^2}{4h_1}\, c_n \log n\}$$

where $c_n = \frac{n}{CD_m \log n}$. Here $h_0, h_1\, C$ are generic constants. Baraud gives a proof for the case $Var(Y|X) =$ constant, but this is immaterial since only functions of $\underset{\sim}{X}$ are involved in $\tilde{\rho_m}$. Therefore,

$$\begin{aligned}
& E_P(\hat{F}_m - F_P)^2 1(\rho_m \leq \rho_0) \\
\leq \; & 2\rho_0^2 E_P\{E_n(\hat{F}_m - F_m)^2 + E_n(F_m - F_P)^2\} \\
\leq \; & C(\frac{D_m}{n} + \|F_m - F_P\|^2)
\end{aligned} \tag{42}$$

On the other hand,

$$\begin{aligned}
E_P(\hat{F}_m - F_P)^2 1(\rho_m > \rho_0) &\leq 2P[\rho_m > \rho_0] \\
&= CD_m^2 \exp\{-AC_n \log n\}
\end{aligned} \tag{43}$$

Combining (42) and (43) we obtain Lemma 11 for $\Delta_{m,n} = 0$, $\hat{F}_m = \widetilde{F}_m$. Putting in $\widetilde{F}_m$ we add a term $CE_P(\hat{F}_m - \widetilde{F}_m)^2$. We now apply Lemma 10 c) and the argument we used to obtain (42) and (43). ■

Proof of Theorem 8: Note that we are limited to rates of convergence which are slower than $n^{-\frac{1}{2}}$. This comes from the combination of R1(i) and bounding the operator by the l_∞ norm of the Gram matrix. It is not clear how either of these conditions can be relaxed.

We need only check that if the $\{\widetilde{F}_m\}$ are the θ_m of Theorem 6 then the conditions of that theorem are satisfied. By construction, $\|\widetilde{F}_m\|_\infty \leq 1$, $B_n = \frac{n}{\log n}$. By Lemma 11 and (R3),

$$r_n \leq C_1 \frac{D_m}{n} + C_2 D_m^{-B} \tag{44}$$

and the right hand side of (44) is bounded by $n^{-(\frac{\beta}{\beta+1})}$. ■

References

P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: A large sample study. *Ann. Stat.* 10:1100–1120, 1982.

Y. Baraud. Model selection for regression on a random design.*Tech. Report*, U. Paris Sud, 2001.

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection under penalization. *Prob. Theory and Related Fields*, 113:301–413, 1999.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Tech. Report* 638, Department of Statistics, University of California at Berkeley, 2003.

P. J. Bickel and P. W. Millar. Uniform convergence of probability measures on classes of functions. Statistica Sinica 2:1-15, 1992.

P. J. Bickel and Y. Ritov. The golden chain. A comment. *Ann. Statist.*, 32:91–96, 2003.

L. Breiman. Arcing classifiers (with discussion). *Ann. Statist.* 26:801–849, 1998.

L. Breiman. Prediction games and arcing algorithms. *Neural Computation* 11:1493-1517, 1999.

L. Breiman. Some infinity theory for predictor ensembles *Technical Report* U.C. Berkeley, 2000.

P. Bühlmann. Consistency for L_2 boosting and matching pursuit with trees and tree type base functions. *Technical Report* ETH Zürich, 2002.

P. Bühlmann and B. Yu. Boosting the L_2 loss: regression and classification. *J. of Amer. Statist. Assoc.*, 98:324–339, 2003

D. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia (with discussion). *J. Roy. Statist. Soc. Ser. B* 57:371–394, 1995.

J. H. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion). *Ann. Statist.* 28:337–407, 2000.

Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation* 121:256–285, 1995.

Y, Freund and R. E. Schapire. Experiments with a new boosting algorithm. *Machine Learning: Proc. 13th International Conference*, 148–156. Morgan Kauffman, San Francisco, 1996.

G. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution Free Theory of Nonparametric Regression*. Springer, New York, 2002.

W. Jiang. Process consistency for ADABOOST. Technical Report 00-05, Dept. of Statistics, Northwestern University, 2002.

Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data . *J. of Amer. Statist. Assoc.*, 99:67–81, 2002.

D. G. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley Publishing Company, Reading, 1984.

G. Lugosi and N. Vayatis. On the Bayes-risk consistency of boosting methods. *Ann. Statist.* 32:30–55, 2004.

S. Mallat and Z. Zhang. Matching pursuit with time frequency dictionaries. *IEEE Transactions on Signal Processing* 41:3397–3415, 1993.

E. Mammen and A. Tsybakov. Smooth discrimination analysis. *Ann. Statist.* 27:1808–1829, 1999.

S. Mannor, R. Meir, and T. Zhang. Greedy algorithms for classification—consistency, convergence rates and adaptivity. J. of Machine Learning Research 4:713–742, 2004.

L. Mason, P. Bartlett, J. Baxter, and M. Frean. Functional gradient techniques for combining hypotheses. In Schölkopg, Smola, A., Bartlett, P., and Schurmans, D. (edts.) *Advances in Large Margin Classifiers*, MIT Press, Boston, 2000.

R. E. Schapire. The strength of weak learnability. *Machine Learning* 5:197–227, 1990.

R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence related predictions. *Machine Learning*, 37:297–336, 1999.

A, Tsybakov. Optimal aggregation of classifiers in statistical learning. *Technical Report*, U. of Paris IV, 2001.

A. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

Y. Yang. Minimax nonparametric classification – Part I Rates of convergence, Part II Model selection, *IEEE Trans. Inf. Theory* 45:2271–2292, 1999.

T. Zhang and B. Yu. Boosting with early stopping: convergence and consistency. Tech Report 635, Stat Dept, UCB, 2003.

T. Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32:56–134, 2004.

The Annals of Statistics
2009, Vol. 37, No. 4, 1705–1732
DOI: 10.1214/08-AOS620

SIMULTANEOUS ANALYSIS OF LASSO AND DANTZIG SELECTOR[1]

BY PETER J. BICKEL, YA'ACOV RITOV AND ALEXANDRE B. TSYBAKOV

University of California at Berkeley, The Hebrew University and Université Paris VI and CREST

We show that, under a sparsity scenario, the Lasso estimator and the Dantzig selector exhibit similar behavior. For both methods, we derive, in parallel, oracle inequalities for the prediction risk in the general nonparametric regression model, as well as bounds on the ℓ_p estimation loss for $1 \le p \le 2$ in the linear model when the number of variables can be much larger than the sample size.

1. Introduction. During the last few years, a great deal of attention has been focused on the ℓ_1 penalized least squares (Lasso) estimator of parameters in high-dimensional linear regression when the number of variables can be much larger than the sample size [8, 9, 11, 17, 18, 20–22, 26] and [27]. Quite recently, Candes and Tao [7] have proposed a new estimate for such linear models, the Dantzig selector, for which they establish optimal ℓ_2 rate properties under a sparsity scenario; that is, when the number of nonzero components of the true vector of parameters is small.

Lasso estimators have also been studied in the nonparametric regression setup [2–4, 12, 13, 19] and [5]. In particular, Bunea, Tsybakov and Wegkamp [2–5] obtain sparsity oracle inequalities for the prediction loss in this context and point out the implications for minimax estimation in classical nonparametric regression settings, as well as for the problem of aggregation of estimators. An analog of Lasso for density estimation with similar properties (SPADES) is proposed in [6]. Modified versions of Lasso estimators (nonquadratic terms and/or penalties slightly different from ℓ_1) for nonparametric regression with random design are suggested and studied under prediction loss in [14] and [25]. Sparsity oracle inequalities for the Dantzig selector with random design are obtained in [15]. In linear fixed design regression, Meinshausen and Yu [18] establish a bound on the ℓ_2 loss for the coefficients of Lasso that is quite different from the bound on the same loss for the Dantzig selector proven in [7].

The main message of this paper is that, under a sparsity scenario, the Lasso and the Dantzig selector exhibit similar behavior, both for linear regression and

Received August 2007; revised April 2008.
[1]Supported in part by NSF Grant DMS-06-05236, ISF grant, France-Berkeley Fund, the Grant ANR-06-BLAN-0194 and the European Network of Excellence PASCAL.
AMS 2000 subject classifications. Primary 60K35, 62G08; secondary 62C20, 62G05, 62G20.
Key words and phrases. Linear models, model selection, nonparametric statistics.

for nonparametric regression models, for ℓ_2 prediction loss and for ℓ_p loss in the coefficients for $1 \le p \le 2$. All the results of the paper are nonasymptotic.

Let us specialize to the case of linear regression with many covariates, $\mathbf{y} = X\beta + \mathbf{w}$, where X is the $n \times M$ deterministic design matrix, with M possibly much larger than n, and $\mathbf{w}$ is a vector of i.i.d. standard normal random variables. This is the situation considered most recently by Candes and Tao [7] and Meinshausen and Yu [18]. Here, sparsity specifies that the high-dimensional vector β has coefficients that are mostly 0.

We develop general tools to study these two estimators in parallel. For the fixed design Gaussian regression model, we recover, as particular cases, sparsity oracle inequalities for the Lasso, as in Bunea, Tsybakov and Wegkamp [4], and ℓ_2 bounds for the coefficients of Dantzig selector, as in Candes and Tao [7]. This is obtained as a consequence of our more general results, which are the following:

- In the nonparametric regression model, we prove sparsity oracle inequalities for the Dantzig selector; that is, bounds on the prediction loss in terms of the best possible (oracle) approximation under the sparsity constraint.
- Similar sparsity oracle inequalities are proved for the Lasso in the nonparametric regression model, and this is done under more general assumptions on the design matrix than in [4].
- We prove that, for nonparametric regression, the Lasso and the Dantzig selector are approximately equivalent in terms of the prediction loss.
- We develop geometrical assumptions that are considerably weaker than those of Candes and Tao [7] for the Dantzig selector and Bunea, Tsybakov and Wegkamp [4] for the Lasso. In the context of linear regression where the number of variables is possibly much larger than the sample size, these assumptions imply the result of [7] for the ℓ_2 loss and generalize it to ℓ_p loss $1 \le p \le 2$ and to prediction loss. Our bounds for the Lasso differ from those for Dantzig selector only in numerical constants.

We begin, in the next section, by defining the Lasso and Dantzig procedures and the notation. In Section 3, we present our key geometric assumptions. Some sufficient conditions for these assumptions are given in Section 4, where they are also compared to those of [7] and [18], as well as to ones appearing in [4] and [5]. We note a weakness of our assumptions, and, hence, of those in the papers we cited, and we discuss a way of slightly remedying them. Sections 5 and 6 give some equivalence results and sparsity oracle inequalities for the Lasso and Dantzig estimators in the general nonparametric regression model. Section 7 focuses on the linear regression model and includes a final discussion. Two important technical lemmas are given in Appendix B as well as most of the proofs.

2. Definitions and notation. Let $(Z_1, Y_1), \dots, (Z_n, Y_n)$ be a sample of independent random pairs with

$$Y_i = f(Z_i) + W_i, \qquad i = 1, \dots, n,$$

where $f : \mathcal{Z} \to \mathbb{R}$ is an unknown regression function to be estimated, $\mathcal{Z}$ is a Borel subset of $\mathbb{R}^d$, the Z_i's are fixed elements in $\mathcal{Z}$ and the regression errors W_i are Gaussian. Let $\mathcal{F}_M = \{f_1, \dots, f_M\}$ be a finite dictionary of functions $f_j : \mathcal{Z} \to \mathbb{R}$, $j = 1, \dots, M$. We assume throughout that $M \geq 2$.

Depending on the statistical targets, the dictionary $\mathcal{F}_M$ can contain qualitatively different parts. For instance, it can be a collection of basis functions used to approximate f in the nonparametric regression model (e.g., wavelets, splines with fixed knots, step functions). Another example is related to the aggregation problem, where the f_j are estimators arising from M different methods. They can also correspond to M different values of the tuning parameter of the same method. Without much loss of generality, these estimators f_j are treated as fixed functions. The results are viewed as being conditioned on the sample that the f_j are based on.

The selection of the dictionary can be very important to make the estimation of f possible. We assume implicitly that f can be well approximated by a member of the span of $\mathcal{F}_M$. However, this is not enough. In this paper, we have in mind the situation where $M \gg n$, and f can be estimated reasonably only because it can approximated by a linear combination of a small number of members of $\mathcal{F}_M$, or, in other words, it has a sparse approximation in the span of $\mathcal{F}_M$. But, when sparsity is an issue, equivalent bases can have different properties. A function that has a sparse representation in one basis may not have it in another, even if both of them span the same linear space.

Consider the matrix $X = (f_j(Z_i))_{i,j}$, $i = 1, \dots, n$, $j = 1, \dots, M$ and the vectors $\mathbf{y} = (Y_1, \dots, Y_n)^{\mathrm{T}}$, $\mathbf{f} = (f(Z_1), \dots, f(Z_n))^{\mathrm{T}}$, $\mathbf{w} = (W_1, \dots, W_n)^{\mathrm{T}}$. With the notation

$$\mathbf{y} = \mathbf{f} + \mathbf{w},$$

we will write $|x|_p$ for the ℓ_p norm of $x \in \mathbb{R}^M$, $1 \leq p \leq \infty$. The notation $\|\cdot\|_n$ stands for the empirical norm

$$\|g\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n g^2(Z_i)}$$

for any $g : \mathcal{Z} \to \mathbb{R}$. We suppose that $\|f_j\|_n \neq 0$, $j = 1, \dots, M$. Set

$$f_{\max} = \max_{1 \leq j \leq M} \|f_j\|_n, \qquad f_{\min} = \min_{1 \leq j \leq M} \|f_j\|_n.$$

For any $\beta = (\beta_1, \dots, \beta_M) \in \mathbb{R}^M$, define $f_\beta = \sum_{j=1}^M \beta_j f_j$ or, explicitly, $f_\beta(z) = \sum_{j=1}^M \beta_j f_j(z)$ and $\mathbf{f}_\beta = X\beta$. The estimates we consider are all of the form $f_{\tilde{\beta}}(\cdot)$, where $\tilde{\beta}$ is data determined. Since we consider mainly sparse vectors $\tilde{\beta}$, it will be convenient to define the following. Let

$$\mathcal{M}(\beta) = \sum_{j=1}^M I_{\{\beta_j \neq 0\}} = |J(\beta)|$$

denote the number of nonzero coordinates of β, where $I_{\{\cdot\}}$ denotes the indicator function $J(\beta) = \{j \in \{1, \dots, M\} : \beta_j \neq 0\}$ and $|J|$ denotes the cardinality of J. The value $\mathcal{M}(\beta)$ characterizes the *sparsity* of the vector β. The smaller $\mathcal{M}(\beta)$, the "sparser" β. For a vector $\boldsymbol{\delta} \in \mathbb{R}^M$ and a subset $J \subset \{1, \dots, M\}$, we denote by $\boldsymbol{\delta}_J$ the vector in $\mathbb{R}^M$ that has the same coordinates as $\boldsymbol{\delta}$ on J and zero coordinates on the complement J^c of J.

Introduce the residual sum of squares

$$\widehat{S}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - f_\beta(Z_i)\}^2$$

for all $\beta \in \mathbb{R}^M$. Define the Lasso solution $\widehat{\beta}_L = (\widehat{\beta}_{1,L}, \dots, \widehat{\beta}_{M,L})$ by

$$\widehat{\beta}_L = \arg\min_{\beta \in \mathbb{R}^M} \left\{ \widehat{S}(\beta) + 2r \sum_{j=1}^{M} \|f_j\|_n |\beta_j| \right\}, \tag{2.1}$$

where $r > 0$ is some tuning constant, and introduce the corresponding Lasso estimator

$$\widehat{f}_L(x) = f_{\widehat{\beta}_L}(x) = \sum_{j=1}^{M} \widehat{\beta}_{j,L} f_j(z). \tag{2.2}$$

The criterion in (2.1) is convex in β, so that standard convex optimization procedures can be used to compute $\widehat{\beta}_L$. We refer to [9, 10, 20, 21, 24] and [16] for detailed discussion of these optimization problems and fast algorithms.

A necessary and sufficient condition of the minimizer in (2.1) is that 0 belongs to the subdifferential of the convex function $\beta \mapsto n^{-1}|y - X\beta|_2^2 + 2r|D^{1/2}\beta|_1$. This implies that the Lasso selector $\widehat{\beta}_L$ satisfies the constraint

$$\left| \frac{1}{n} D^{-1/2} X^{\mathrm{T}} (y - X\widehat{\beta}_L) \right|_\infty \le r, \tag{2.3}$$

where D is the diagonal matrix

$$D = \operatorname{diag}\{\|f_1\|_n^2, \dots, \|f_M\|_n^2\}.$$

More generally, we will say that $\beta \in \mathbb{R}^M$ satisfies the Dantzig constraint if β belongs to the set

$$\left\{ \beta \in \mathbb{R}^M : \left| \frac{1}{n} D^{-1/2} X^{\mathrm{T}} (y - X\beta) \right|_\infty \le r \right\}.$$

The Dantzig estimator of the regression function f is based on a particular solution of (2.3), the Dantzig selector $\widehat{\beta}_D$, which is defined as a vector having the smallest ℓ_1 norm among all β satisfying the Dantzig constraint

$$\widehat{\beta}_D = \arg\min \left\{ |\beta|_1 : \left| \frac{1}{n} D^{-1/2} X^{\mathrm{T}} (y - X\beta) \right|_\infty \le r \right\}. \tag{2.4}$$

The Dantzig estimator is defined by

$$\widehat{f}_D(z) = f_{\widehat{\beta}_D}(z) = \sum_{j=1}^{M} \widehat{\beta}_{j,D} f_j(z), \tag{2.5}$$

where $\widehat{\beta}_D = (\widehat{\beta}_{1,D}, \dots, \widehat{\beta}_{M,D})$ is the Dantzig selector. By the definition of Dantzig selector, we have $|\widehat{\beta}_D|_1 \le |\widehat{\beta}_L|_1$.

The Dantzig selector is computationally feasible, since it reduces to a linear programming problem [7].

Finally, for any $n \ge 1$, $M \ge 2$, we consider the Gram matrix

$$\Psi_n = \frac{1}{n} X^{\mathrm{T}} X = \left(\frac{1}{n} \sum_{i=1}^{n} f_j(Z_i) f_{j'}(Z_i) \right)_{1 \le j, j' \le M},$$

and let $\phi_{\max}$ denote the maximal eigenvalue of Ψ_n.

3. Restricted eigenvalue assumptions. We now introduce the key assumptions on the Gram matrix that are needed to guarantee nice statistical properties of the Lasso and Dantzig selectors. Under the sparsity scenario, we are typically interested in the case where $M > n$, and even $M \gg n$. Then, the matrix Ψ_n is degenerate, which can be written as

$$\min_{\delta \in \mathbb{R}^M : \delta \ne 0} \frac{(\boldsymbol{\delta}^{\mathrm{T}} \Psi_n \boldsymbol{\delta})^{1/2}}{|\boldsymbol{\delta}|_2} \equiv \min_{\delta \in \mathbb{R}^M : \delta \ne 0} \frac{|X\boldsymbol{\delta}|_2}{\sqrt{n}|\boldsymbol{\delta}|_2} = 0.$$

Clearly, ordinary least squares does not work in this case, since it requires positive definiteness of Ψ_n; that is,

$$\min_{\delta \in \mathbb{R}^M : \delta \ne 0} \frac{|X\boldsymbol{\delta}|_2}{\sqrt{n}|\boldsymbol{\delta}|_2} > 0. \tag{3.1}$$

It turns out that the Lasso and Dantzig selector require much weaker assumptions. The minimum in (3.1) can be replaced by the minimum over a restricted set of vectors, and the norm $|\boldsymbol{\delta}|_2$ in the denominator of the condition can be replaced by the ℓ_2 norm of only a part of $\boldsymbol{\delta}$.

One of the properties of both the Lasso and the Dantzig selectors is that, for the linear regression model, the residuals $\boldsymbol{\delta} = \hat{\beta}_L - \beta$ and $\boldsymbol{\delta} = \hat{\beta}_D - \beta$ satisfy, with probability close to 1,

$$|\boldsymbol{\delta}_{J_0^c}|_1 \le c_0 |\boldsymbol{\delta}_{J_0}|_1, \tag{3.2}$$

where $J_0 = J(\beta)$ is the set of nonzero coefficients of the true parameter β of the model. For the linear regression model, the vector of Dantzig residuals $\boldsymbol{\delta}$ satisfies (3.2) with probability close 1 if $c_0 = 1$ and M is large [cf. (B.9) and the fact that β of the model satisfies the Dantzig constraint with probability close to 1 if M is large]. A similar inequality holds for the vector of Lasso residuals $\boldsymbol{\delta} = \widehat{\beta}_L - \beta$, but this time with $c_0 = 3$ [cf. Corollary B.2].

Now, for example, consider the case where the elements of the Gram matrix Ψ_n are close to those of a positive definite $(M \times M)$-matrix Ψ. Denote, by $\varepsilon_n \triangleq \max_{i,j} |(\Psi_n - \Psi)_{i,j}|$, the maximal difference between the elements of the two matrices. Then, for any $\boldsymbol{\delta}$ satisfying (3.2), we get

$$
\begin{aligned}
\frac{\boldsymbol{\delta}^{\mathrm{T}}\Psi_n\boldsymbol{\delta}}{|\boldsymbol{\delta}|_2^2} &= \frac{\boldsymbol{\delta}^{\mathrm{T}}\Psi\boldsymbol{\delta} + \boldsymbol{\delta}^{\mathrm{T}}(\Psi_n - \Psi)\boldsymbol{\delta}}{|\boldsymbol{\delta}|_2^2} \\
&\geq \frac{\boldsymbol{\delta}^{\mathrm{T}}\Psi\boldsymbol{\delta}}{|\boldsymbol{\delta}|_2^2} - \frac{\varepsilon_n|\boldsymbol{\delta}|_1^2}{|\boldsymbol{\delta}|_2^2} \\
&\geq \frac{\boldsymbol{\delta}^{\mathrm{T}}\Psi\boldsymbol{\delta}}{|\boldsymbol{\delta}|_2^2} - \varepsilon_n\left(\frac{(1+c_0)|\boldsymbol{\delta}_{J_0}|_1}{|\boldsymbol{\delta}_{J_0}|_2}\right)^2 \\
&\geq \frac{\boldsymbol{\delta}^{\mathrm{T}}\Psi\boldsymbol{\delta}}{|\boldsymbol{\delta}|_2^2} - \varepsilon_n(1+c_0)^2|J_0|.
\end{aligned} \tag{3.3}
$$

Thus, for $\boldsymbol{\delta}$ satisfying (3.2), which are the vectors that we have in mind, and for $\varepsilon_n|J_0|$ small enough, the LHS of (3.3) is bounded away from 0. This means that we have a kind of "restricted" positive definiteness, which is valid only for the vectors satisfying (3.2). This suggests the following conditions, which will suffice for the main argument of the paper. We refer to these conditions as *restricted eigenvalue* (RE) assumptions.

ASSUMPTION RE(s, c_0). For some integer s such that $1 \leq s \leq M$ and a positive number c_0, the following condition holds:

$$
\kappa(s, c_0) \triangleq \min_{\substack{J_0 \subseteq \{1,\dots,M\},\\ |J_0| \leq s}} \quad \min_{\substack{\boldsymbol{\delta} \neq 0,\\ |\boldsymbol{\delta}_{J_0^c}|_1 \leq c_0|\boldsymbol{\delta}_{J_0}|_1}} \frac{|X\boldsymbol{\delta}|_2}{\sqrt{n}|\boldsymbol{\delta}_{J_0}|_2} > 0.
$$

The integer s here plays the role of an upper bound on the sparsity $\mathcal{M}(\beta)$ of a vector of coefficients β.

Note that, if Assumption RE(s, c_0) is satisfied with $c_0 \geq 1$, then

$$
\min\{|X\boldsymbol{\delta}|_2 : \mathcal{M}(\boldsymbol{\delta}) \leq 2s, \boldsymbol{\delta} \neq 0\} > 0.
$$

In other words, the square submatrices of size $\leq 2s$ of the Gram matrix are necessarily positive definite. Indeed, suppose that, for some $\boldsymbol{\delta} \neq 0$, we have simultaneously $\mathcal{M}(\boldsymbol{\delta}) \leq 2s$ and $X\boldsymbol{\delta} = 0$. Partition $J(\boldsymbol{\delta})$ in two sets $J(\boldsymbol{\delta}) = I_0 \cup I_1$, such that $|I_i| \leq s$, $i = 0, 1$. Without loss of generality, suppose that $|\boldsymbol{\delta}_{I_1}|_1 \leq |\boldsymbol{\delta}_{I_0}|_1$. Since, clearly, $|\boldsymbol{\delta}_{I_1}|_1 = |\boldsymbol{\delta}_{I_0^c}|_1$ and $c_0 \geq 1$, we have $|\boldsymbol{\delta}_{I_0^c}|_1 \leq c_0|\boldsymbol{\delta}_{I_0}|_1$. Hence, $\kappa(s, c_0) = 0$, a contradiction.

To introduce the second assumption, we need some notation. For integers s, m such that $1 \le s \le M/2$ and $m \ge s$, $s + m \le M$, a vector $\boldsymbol{\delta} \in \mathbb{R}^M$ and a set of indices $J_0 \subseteq \{1, \dots, M\}$ with $|J_0| \le s$; denote by J_1 the subset of $\{1, \dots, M\}$ corresponding to the m largest in absolute value coordinates of $\boldsymbol{\delta}$ outside of J_0, and define $J_{01} \stackrel{\triangle}{=} J_0 \cup J_1$. Clearly, J_1 and J_{01} depend on m, but we do not indicate this in our notation for the sake of brevity.

ASSUMPTION RE(s, m, c_0).

$$\kappa(s, m, c_0) \stackrel{\triangle}{=} \min_{\substack{J_0 \subseteq \{1,\dots,M\},\\ |J_0| \le s}} \quad \min_{\substack{\boldsymbol{\delta} \ne 0,\\ |\boldsymbol{\delta}_{J_0^c}|_1 \le c_0 |\boldsymbol{\delta}_{J_0}|_1}} \frac{|X\boldsymbol{\delta}|_2}{\sqrt{n}|\boldsymbol{\delta}_{J_{01}}|_2} > 0.$$

Note that the only difference between the two assumptions is in the denominators, and $\kappa(s, m, c_0) \le \kappa(s, c_0)$. As written, for fixed n, the two assumptions are equivalent. However, asymptotically for large n, Assumption RE(s, c_0) is less restrictive than RE(s, m, c_0), since the ratio $\kappa(s, m, c_0)/\kappa(s, c_0)$ may tend to 0 if s and m depend on n. For our bounds on the prediction loss and on the ℓ_1 loss of the Lasso and Dantzig estimators, we will only need Assumption RE(s, c_0). Assumption RE(s, m, c_0) will be required exclusively for the bounds on the ℓ_p loss with $1 < p \le 2$.

Note also that Assumptions RE(s', c_0) and RE(s', m, c_0) imply Assumptions RE(s, c_0) and RE(s, m, c_0), respectively, if $s' > s$.

4. Discussion of the RE assumptions. There exist several simple sufficient conditions for Assumptions RE(s, c_0) and RE(s, m, c_0) to hold. Here, we discuss some of them.

For a real number $1 \le u \le M$, we introduce the following quantities that we will call *restricted eigenvalues*:

$$\phi_{\min}(u) = \min_{x \in \mathbb{R}^M : 1 \le \mathcal{M}(x) \le u} \frac{x^{\mathrm{T}} \Psi_n x}{|x|_2^2},$$

$$\phi_{\max}(u) = \max_{x \in \mathbb{R}^M : 1 \le \mathcal{M}(x) \le u} \frac{x^{\mathrm{T}} \Psi_n x}{|x|_2^2}.$$

Denote by X_J the $n \times |J|$ submatrix of X obtained by removing from X the columns that do not correspond to the indices in J, and, for $1 \le m_1, m_2 \le M$, introduce the following quantities called *restricted correlations*:

$$\theta_{m_1,m_2} = \max\left\{\frac{c_1^{\mathrm{T}} X_{I_1}^{\mathrm{T}} X_{I_2} c_2}{n|c_1|_2|c_2|_2} : I_1 \cap I_2 = \varnothing, |I_i| \le m_i, c_i \in \mathbb{R}^{I_i} \setminus \{0\}, i = 1, 2\right\}.$$

In Lemma 4.1, below, we show that a sufficient condition for RE(s, c_0) and RE(s, s, c_0) to hold is given, for example, by the following assumption on the Gram matrix.

ASSUMPTION 1. Assume that

$$\phi_{\min}(2s) > c_0\theta_{s,2s}$$

for some integer $1 \le s \le M/2$ and a constant $c_0 > 0$.

This condition with $c_0 = 1$ appeared in [7], in connection with the Dantzig selector. Assumption 1 is more general, in that we can have an arbitrary constant $c_0 > 0$ that will allow us to cover not only the Dantzig selector but also the Lasso estimators and to prove oracle inequalities for the prediction loss when the model is nonparametric.

Our second sufficient condition for RE(s, c_0) and RE(s, m, c_0) does not need bounds on correlations. Only bounds on the minimal and maximal eigenvalues of "small" submatrices of the Gram matrix Ψ_n are involved.

ASSUMPTION 2. Assume that

$$m\phi_{\min}(s+m) > c_0^2 s\phi_{\max}(m)$$

for some integers s, m, such that $1 \le s \le M/2$, $m \ge s$ and $s + m \le M$, and a constant $c_0 > 0$.

Assumption 2 can be viewed as a weakening of the condition on $\phi_{\min}$ in [18]. Indeed, taking $s + m = s\log n$ (we assume, without loss of generality, that $s\log n$ is an integer and $n > 3$) and assuming that $\phi_{\max}(\cdot)$ is uniformly bounded by a constant, we get that Assumption 2 is equivalent to

$$\phi_{\min}(s\log n) > c/\log n,$$

where $c > 0$ is a constant. The corresponding, slightly stronger, assumption in [18] is stated in asymptotic form, for $s = s_n \to \infty$, as

$$\liminf_n \phi_{\min}(s_n \log n) > 0.$$

The following two constants are useful when Assumptions 1 and 2 are considered:

$$\kappa_1(s, c_0) = \sqrt{\phi_{\min}(2s)}\left(1 - \frac{c_0\theta_{s,2s}}{\phi_{\min}(2s)}\right)$$

and

$$\kappa_2(s, m, c_0) = \sqrt{\phi_{\min}(s+m)}\left(1 - c_0\sqrt{\frac{s\phi_{\max}(m)}{m\phi_{\min}(s+m)}}\right).$$

The next lemma shows that if Assumptions 1 or 2 are satisfied, then the quadratic form $x^{\mathrm{T}}\Psi_n x$ is positive definite on some restricted sets of vectors x. The construction of the lemma is inspired by Candes and Tao [7] and covers, in particular, the corresponding result in [7].

LEMMA 4.1. *Fix an integer* $1 \le s \le M/2$ *and a constant* $c_0 > 0$.

(i) *Let Assumption* 1 *be satisfied. Then, Assumptions* RE(s, c_0) *and* RE(s, s, c_0) *hold with* $\kappa(s, c_0) = \kappa(s, s, c_0) = \kappa_1(s, c_0)$. *Moreover, for any subset* J_0 *of* $\{1, \dots, M\}$, *with cardinality* $|J_0| \le s$, *and any* $\boldsymbol{\delta} \in \mathbb{R}^M$ *such that*

$$|\boldsymbol{\delta}_{J_0^c}|_1 \le c_0 |\boldsymbol{\delta}_{J_0}|_1, \tag{4.1}$$

we have

$$\frac{1}{\sqrt{n}} |P_{01} X \boldsymbol{\delta}|_2 \ge \kappa_1(s, c_0) |\boldsymbol{\delta}_{J_{01}}|_2,$$

where P_{01} *is the projector in* $\mathbb{R}^M$ *on the linear span of the columns of* $X_{J_{01}}$.

(ii) *Let Assumption* 2 *be satisfied. Then, Assumptions* RE(s, c_0) *and* RE(s, m, c_0) *hold with* $\kappa(s, c_0) = \kappa(s, m, c_0) = \kappa_2(s, m, c_0)$. *Moreover, for any subset* J_0 *of* $\{1, \dots, M\}$, *with cardinality* $|J_0| \le s$, *and any* $\boldsymbol{\delta} \in \mathbb{R}^M$ *such that* (4.1) *holds, we have*

$$\frac{1}{\sqrt{n}} |P_{01} X \boldsymbol{\delta}|_2 \ge \kappa_2(s, m, c_0) |\boldsymbol{\delta}_{J_{01}}|_2.$$

The proof of the lemma is given in Appendix A.

There exist other sufficient conditions for Assumptions RE(s, c_0) and RE(s, m, c_0) to hold. We mention here three of them implying Assumption RE(s, c_0). The first one is the following [1].

ASSUMPTION 3. For an integer s such that $1 \le s \le M$, we have

$$\phi_{\min}(s) > 2 c_0 \theta_{s,1} \sqrt{s},$$

where $c_0 > 0$ is a constant.

To argue that Assumption 3 implies RE(s, c_0), it suffices to remark that

$$\begin{aligned} \frac{1}{n} |X\boldsymbol{\delta}|_2^2 &\ge \frac{1}{n} \boldsymbol{\delta}_{J_0}^{\mathrm{T}} X^{\mathrm{T}} X \boldsymbol{\delta}_{J_0} - \frac{2}{n} |\boldsymbol{\delta}_{J_0}^{\mathrm{T}} X^{\mathrm{T}} X \boldsymbol{\delta}_{J_0^c}| \\ &\ge \phi_{\min}(s) |\boldsymbol{\delta}_{J_0}|_2^2 - \frac{2}{n} |\boldsymbol{\delta}_{J_0}^{\mathrm{T}} X^{\mathrm{T}} X \boldsymbol{\delta}_{J_0^c}| \end{aligned}$$

and, if (4.1) holds,

$$\begin{aligned} |\boldsymbol{\delta}_{J_0}^{\mathrm{T}} X^{\mathrm{T}} X \boldsymbol{\delta}_{J_0^c}|/n &\le |\boldsymbol{\delta}_{J_0^c}|_1 \max_{j \in J_0^c} |\boldsymbol{\delta}_{J_0}^{\mathrm{T}} X^{\mathrm{T}} \mathbf{x}_{(j)}|/n \\ &\le \theta_{s,1} |\boldsymbol{\delta}_{J_0^c}|_1 |\boldsymbol{\delta}_{J_0}|_2 \\ &\le c_0 \theta_{s,1} \sqrt{s} |\boldsymbol{\delta}_{J_0}|_2^2. \end{aligned}$$

Another type of assumption related to "mutual coherence" [8] is discussed in connection to Lasso in [4, 5]. We state it in two different forms, which are given below.

ASSUMPTION 4. For an integer s such that $1 \le s \le M$, we have

$$\phi_{\min}(s) > 2c_0\theta_{1,1}s,$$

where $c_0 > 0$ is a constant.

It is easy to see that Assumption 4 implies RE(s, c_0). Indeed, if (4.1) holds,

$$\begin{aligned}\frac{1}{n}|X\delta|_2^2 &\ge \frac{1}{n}\delta_{J_0}^{\mathrm{T}}X^{\mathrm{T}}X\delta_{J_0} - 2\theta_{1,1}|\delta_{J_0^c}|_1|\delta_{J_0}|_1\\ &\ge \phi_{\min}(s)|\delta_{J_0}|_2^2 - 2c_0\theta_{1,1}|\delta_{J_0}|_1^2 \qquad (4.2)\\ &\ge (\phi_{\min}(s) - 2c_0\theta_{1,1}s)|\delta_{J_0}|_2^2.\end{aligned}$$

If all the diagonal elements of matrix $X^{\mathrm{T}}X/n$ are equal to 1 (and thus $\theta_{1,1}$ coincides with the mutual coherence [8]), then a simple sufficient condition for Assumption RE(s, c_0) to hold is stated as follows.

ASSUMPTION 5. All the diagonal elements of the Gram matrix Ψ_n are equal to 1, and for an integer s, such that $1 \le s \le M$, we have

$$\theta_{1,1} < \frac{1}{(1+2c_0)s}, \qquad (4.3)$$

where $c_0 > 0$ is a constant.

In fact, separating the diagonal and off-diagonal terms of the quadratic form, we get

$$\delta_{J_0}^{\mathrm{T}}X^{\mathrm{T}}X\delta_{J_0}/n \ge |\delta_{J_0}|_2^2 - \theta_{1,1}|\delta_{J_0}|_1^2 \ge |\delta_{J_0}|_2^2(1 - \theta_{1,1}s).$$

Combining this inequality with (4.2), we see that Assumption RE(s, c_0) is satisfied whenever (4.3) holds.

Unfortunately, Assumption RE(s, c_0) has some weakness. Let, for example, f_j, $j = 1, \dots, 2^m - 1$, be the Haar wavelet basis on $[0, 1]$ ($M = 2^m$), and consider $Z_i = i/n$, $i = 1, \dots, n$. If $M \gg n$, then it is clear that $\phi_{\min}(1) = 0$, since there are functions f_j on the highest resolution level whose supports (of length M^{-1}) contain no points Z_i. So, none of Assumptions 1–4 hold. A less severe, although similar, situation is when we consider step functions $f_j(t) = I_{\{t<j/M\}}$ for $t \in [0, 1]$. It is clear that $\phi_{\min}(2) = O(1/M)$, although sparse representation in this basis is very natural. Intuitively, the problem arises only because we include very high resolution components. Therefore, we may try to restrict the set J_0 in RE(s, c_0) to low resolution components, which is quite reasonable, because the "true" or "interesting" vectors of parameters β are often characterized by such J_0. This idea is formalized in Section 6 (cf. Corollary 6.2, see also a remark after Theorem 7.2 in Section 7).

5. Approximate equivalence. In this section, we prove a type of approximate equivalence between the Lasso and the Dantzig selector. It is expressed as closeness of the prediction losses $\|\widehat{f}_D - f\|_n^2$ and $\|\widehat{f}_L - f\|_n^2$ when the number of nonzero components of the Lasso or the Dantzig selector is small as compared to the sample size.

THEOREM 5.1. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Fix $n \geq 1$, $M \geq 2$. Let Assumption* RE$(s, 1)$ *be satisfied with $1 \leq s \leq M$. Consider the Dantzig estimator $\widehat{f}_D$ defined by* (2.5)–(2.4) *with*

$$r = A\sigma\sqrt{\frac{\log M}{n}},$$

where $A > 2\sqrt{2}$, and consider the Lasso estimator $\widehat{f}_L$ defined by (2.1)–(2.2) *with the same r.*

If $\mathcal{M}(\widehat{\beta}_L) \leq s$, then, with probability at least $1 - M^{1-A^2/8}$, we have

$$\big|\|\widehat{f}_D - f\|_n^2 - \|\widehat{f}_L - f\|_n^2\big| \leq 16A^2 \frac{\mathcal{M}(\widehat{\beta}_L)\sigma^2}{n} \frac{f_{\max}^2}{\kappa^2(s, 1)} \log M. \tag{5.1}$$

Note that the RHS of (5.1) is bounded by a product of three factors (and a numerical constant which, unfortunately, equals at least 128). The first factor $\mathcal{M}(\widehat{\beta}_L)\sigma^2/n \leq s\sigma^2/n$ corresponds to the error rate for prediction in regression with s parameters. The two other factors, $\log M$ and $f_{\max}^2/\kappa^2(s, 1)$, can be regarded as a price to pay for the large number of regressors. If the Gram matrix Ψ_n equals the identity matrix (the white noise model), then there is only the $\log M$ factor. In the general case, there is another factor $f_{\max}^2/\kappa^2(s, 1)$ representing the extent to which the Gram matrix is ill-posed for estimation of sparse vectors.

We also have the following result that we state, for simplicity, under the assumption that $\|f_j\|_n = 1$, $j = 1, \ldots, M$. It gives a bound in the spirit of Theorem 5.1 but with $\mathcal{M}(\widehat{\beta}_D)$ rather than $\mathcal{M}(\widehat{\beta}_L)$ on the right-hand side.

THEOREM 5.2. *Let the assumptions of Theorem* 5.1 *hold, but with* RE$(s, 5)$ *in place of* RE$(s, 1)$, *and let $\|f_j\|_n = 1$, $j = 1, \ldots, M$. If $\mathcal{M}(\widehat{\beta}_D) \leq s$, then, with probability at least $1 - M^{1-A^2/8}$, we have*

$$\|\widehat{f}_L - f\|_n^2 \leq 10\|\widehat{f}_D - f\|_n^2 + 81A^2 \frac{\mathcal{M}(\widehat{\beta}_D)\sigma^2}{n} \frac{\log M}{\kappa^2(s, 5)}. \tag{5.2}$$

REMARK. The approximate equivalence is essentially that of the rates as Theorem 5.1 exhibits. A statement free of $\mathcal{M}(\beta)$ holds for linear regression, see discussion after Theorems 7.2 and 7.3 below.

6. Oracle inequalities for prediction loss. Here, we prove sparsity oracle inequalities for the prediction loss of the Lasso and Dantzig estimators. These inequalities allow us to bound the difference between the prediction errors of the estimators and the best sparse approximation of the regression function (by an oracle that knows the truth but is constrained by sparsity). The results of this section, together with those of Section 5, show that the distance between the prediction losses of the Dantzig and Lasso estimators is of the same order as the distances between them and their oracle approximations.

A general discussion of sparsity oracle inequalities can be found in [23]. Such inequalities have been recently obtained for the Lasso type estimators in a number of settings [2–6, 14] and [25]. In particular, the regression model with fixed design that we study here is considered in [2–4]. The assumptions on the Gram matrix Ψ_n in [2–4] are more restrictive than ours. In those papers, either Ψ_n is positive definite, or a mutual coherence condition similar to (4.3) is imposed.

THEOREM 6.1. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Fix some $\varepsilon > 0$ and integers $n \geq 1$, $M \geq 2$, $1 \leq s \leq M$. Let Assumption* RE$(s, 3+4/\varepsilon)$ *be satisfied. Consider the Lasso estimator $\widehat{f}_L$ defined by* (2.1)–(2.2) *with*

$$r = A\sigma\sqrt{\frac{\log M}{n}}$$

for some $A > 2\sqrt{2}$. Then, with probability at least $1 - M^{1-A^2/8}$, we have

$$\begin{aligned} &\|\widehat{f}_L - f\|_n^2 \\ &\quad \leq (1+\varepsilon) \inf_{\substack{\beta \in \mathbb{R}^M: \\ \mathcal{M}(\beta) \leq s}} \left\{ \|f_\beta - f\|_n^2 + \frac{C(\varepsilon) f_{\max}^2 A^2 \sigma^2}{\kappa^2(s, 3+4/\varepsilon)} \frac{\mathcal{M}(\beta)\log M}{n} \right\}, \end{aligned} \tag{6.1}$$

where $C(\varepsilon) > 0$ is a constant depending only on ε.

We now state, as a corollary, a softer version of Theorem 6.1 that can be used to eliminate the pathologies mentioned at the end of Section 4. For this purpose, we define

$$\mathcal{J}_{s,\gamma,c_0} = \left\{ J_0 \subset \{1, \ldots, M\} : |J_0| \leq s \text{ and } \min_{\substack{\delta \neq 0, \\ |\delta_{J_0^c}|_1 \leq c_0 |\delta_{J_0}|_1}} \frac{|X\delta|_2}{\sqrt{n}|\delta_{J_0}|_2} \geq \gamma \right\},$$

where $\gamma > 0$ is a constant, and set

$$\Lambda_{s,\gamma,c_0} = \{\beta : J(\beta) \in \mathcal{J}_{s,\gamma,c_0}\}.$$

In similar way, we define $\mathcal{J}_{s,\gamma,m,c_0}$ and Λ_{s,γ,m,c_0} corresponding to Assumption RE(s, m, c_0).

COROLLARY 6.2. *Let W_i, s and the Lasso estimator $\widehat{f}_L$ be the same as in Theorem* 6.1. *Then, for all $n \geq 1$, $\varepsilon > 0$, and $\gamma > 0$, with probability at least $1 - M^{1-A^2/8}$ we have*

$$\|\widehat{f}_L - f\|_n^2 \leq (1+\varepsilon) \inf_{\beta \in \bar{\Lambda}_{s,\gamma,\varepsilon}} \left\{ \|f_\beta - f\|_n^2 + \frac{C(\varepsilon) f_{\max}^2 A^2 \sigma^2}{\gamma^2} \left(\frac{\mathcal{M}(\beta) \log M}{n} \right) \right\},$$

where $\bar{\Lambda}_{s,\gamma,\varepsilon} = \{\beta \in \Lambda_{s,\gamma,3+4/\varepsilon} : \mathcal{M}(\beta) \leq s\}$.

To obtain this corollary, it suffices to observe that the proof of Theorem 6.1 goes through if we drop Assumption RE$(s, 3 + 4/\varepsilon)$, but we assume instead that $\beta \in \Lambda_{s,\gamma,3+4/\varepsilon}$, and we replace $\kappa(s, 3 + 4/\varepsilon)$ by γ.

We would like now to get a sparsity oracle inequality similar to that of Theorem 6.1 for the Dantzig estimator $\widehat{f}_D$. We will need a mild additional assumption on f. This is due to the fact that not every $\beta \in \mathbb{R}^M$ obeys the Dantzig constraint; thus, we cannot assure the key relation (B.9) for all $\beta \in \mathbb{R}^M$. One possibility would be to prove inequality as (6.1), where the infimum on the right hand side is taken over β satisfying not only $\mathcal{M}(\beta) \leq s$ but also the Dantzig constraint. However, this seems not to be very intuitive, since we cannot guarantee that the corresponding f_β gives a good approximation of the unknown function f. Therefore, we choose another approach (cf. [5]), in which we consider f satisfying the *weak sparsity* property relative to the dictionary $f_1, \dots, f_M$. That is, we assume that there exist an integer s and constant $C_0 < \infty$ such that the set

$$\Lambda_s = \left\{ \beta \in \mathbb{R}^M : \mathcal{M}(\beta) \leq s, \|f_\beta - f\|_n^2 \leq \frac{C_0 f_{\max}^2 r^2}{\kappa^2(s, 3 + 4/\varepsilon)} \mathcal{M}(\beta) \right\} \tag{6.2}$$

is nonempty. The second inequality in (6.2) says that the "bias" term $\|f_\beta - f\|_n^2$ cannot be much larger than the "variance term" $\sim f_{\max}^2 r^2 \kappa^{-2} \mathcal{M}(\beta)$ [cf. (6.1)]. Weak sparsity is milder than the sparsity property in the usual sense. The latter means that f admits the exact representation $f = f_{\beta^*}$, for some $\beta^* \in \mathbb{R}^M$, with hopefully small $\mathcal{M}(\beta^*) = s$.

PROPOSITION 6.3. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Fix some $\varepsilon > 0$ and integers $n \geq 1$, $M \geq 2$. Let f obey the weak sparsity assumption for some $C_0 < \infty$ and some s such that $1 \leq s \max\{C_1(\varepsilon), 1\} \leq M$, where*

$$C_1(\varepsilon) = 4[(1+\varepsilon)C_0 + C(\varepsilon)] \frac{\phi_{\max} f_{\max}^2}{\kappa^2 f_{\min}^2}$$

and $C(\varepsilon)$ is the constant in Theorem 6.1. *Suppose, further, that Assumption* RE$(s \max\{C_1(\varepsilon), 1\}, 3 + 4/\varepsilon)$ *is satisfied. Consider the Dantzig estimator $\widehat{f}_D$ defined by* (2.5)–(2.4) *with*

$$r = A\sigma \sqrt{\frac{\log M}{n}}$$

and $A > 2\sqrt{2}$. *Then, with probability at least* $1 - M^{1-A^2/8}$, *we have*

$$\|\widehat{f}_D - f\|_n^2 \le (1+\varepsilon) \inf_{\beta\in\mathbb{R}^M: \mathcal{M}(\beta)=s} \|f_\beta - f\|_n^2 + C_2(\varepsilon)\frac{f_{\max}^2 A^2\sigma^2}{\kappa_0^2}\left(\frac{s\log M}{n}\right). \tag{6.3}$$

Here, $C_2(\varepsilon) = 16C_1(\varepsilon) + C(\varepsilon)$ *and* $\kappa_0 = \kappa(\max(C_1(\varepsilon), 1)s, 3 + 4/\varepsilon)$.

Note that the sparsity oracle inequality (6.3) is slightly weaker than the analogous inequality (6.1) for the Lasso. Here, we have $\inf_{\beta\in\mathbb{R}^M:\mathcal{M}(\beta)=s}$ instead of $\inf_{\beta\in\mathbb{R}^M:\mathcal{M}(\beta)\le s}$ in (6.1).

7. Special case. Parametric estimation in linear regression. In this section, we assume that the vector of observations $\mathbf{y} = (Y_1, \dots, Y_n)^{\mathrm{T}}$ is of the form

$$\mathbf{y} = X\beta^* + \mathbf{w}, \tag{7.1}$$

where X is an $n \times M$ deterministic matrix $\beta^* \in \mathbb{R}^M$ and $\mathbf{w} = (W_1, \dots, W_n)^{\mathrm{T}}$.

We consider dimension M that can be of order n and even much larger. Then, β^* is, in general, not uniquely defined. For $M > n$, if (7.1) is satisfied for $\beta^* = \beta_0$, then there exists an affine space $\mathcal{U} = \{\beta^* : X\beta^* = X\beta_0\}$ of vectors satisfying (7.1). The results of this section are valid for any β^* such that (7.1) holds. However, we will suppose that Assumption RE(s, c_0) holds with $c_0 \ge 1$ and that $\mathcal{M}(\beta^*) \le s$. Then, the set $\mathcal{U} \cap \{\beta^* : \mathcal{M}(\beta^*) \le s\}$ reduces to a single element (cf. Remark 2 at the end of this section). In this sense, there is a unique sparse solution of (7.1).

Our goal in this section, unlike that of the previous ones, is to estimate both $X\beta^*$ for the purpose of prediction and β^* itself for purpose of model selection. We will see that meaningful results are obtained when the sparsity index $\mathcal{M}(\beta^*)$ is small.

It will be assumed throughout this section that the diagonal elements of the Gram matrix $\Psi_n = X^{\mathrm{T}}X/n$ are all equal to 1 (this is equivalent to the condition $\|f_j\|_n = 1$, $j = 1, \dots, M$, in the notation of previous sections). Then, the Lasso estimator of β^* in (7.1) is defined by

$$\widehat{\beta}_L = \arg\min_{\beta\in\mathbb{R}^M}\left\{\frac{1}{n}|\mathbf{y} - X\beta|_2^2 + 2r|\beta|_1\right\}. \tag{7.2}$$

The correspondence between the notation here and that of the previous sections is

$$\|f_\beta\|_n^2 = |X\beta|_2^2/n, \qquad \|f_\beta - f\|_n^2 = |X(\beta - \beta^*)|_2^2/n,$$
$$\|\widehat{f}_L - f\|_n^2 = |X(\widehat{\beta}_L - \beta^*)|_2^2/n.$$

The Dantzig selector for linear model (7.1) is defined by

$$\widehat{\beta}_D = \arg\min_{\beta\in\Lambda} |\beta|_1, \tag{7.3}$$

where

$$\Lambda = \left\{ \beta \in \mathbb{R}^M : \left| \frac{1}{n} X^{\mathrm{T}} (\mathbf{y} - X\beta) \right|_\infty \le r \right\}$$

is the set of all β satisfying the Dantzig constraint.

We first get bounds on the rate of convergence of Dantzig selector.

THEOREM 7.1. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$, let all the diagonal elements of the matrix $X^{\mathrm{T}}X/n$ be equal to* 1 *and $\mathcal{M}(\beta^*) \le s$, where $1 \le s \le M$, $n \ge 1$, $M \ge 2$. Let Assumption* RE$(s, 1)$ *be satisfied. Consider the Dantzig selector $\widehat{\beta}_D$ defined by* (7.3) *with*

$$r = A\sigma \sqrt{\frac{\log M}{n}}$$

and $A > \sqrt{2}$. Then, with probability at least $1 - M^{1-A^2/2}$, we have

$$|\widehat{\beta}_D - \beta^*|_1 \le \frac{8A}{\kappa^2(s,1)} \sigma s \sqrt{\frac{\log M}{n}}, \tag{7.4}$$

$$|X(\widehat{\beta}_D - \beta^*)|_2^2 \le \frac{16A^2}{\kappa^2(s,1)} \sigma^2 s \log M. \tag{7.5}$$

If Assumption RE$(s, m, 1)$ *is satisfied, then, with the same probability as above, simultaneously for all $1 < p \le 2$, we have*

$$|\widehat{\beta}_D - \beta^*|_p^p \le 2^{p-1} 8 \left\{ 1 + \sqrt{\frac{s}{m}} \right\}^{2(p-1)} s \left(\frac{A\sigma}{\kappa^2(s,m,1)} \sqrt{\frac{\log M}{n}} \right)^p. \tag{7.6}$$

Note that, since $s \le m$, the factor in curly brackets in (7.6) is bounded by a constant independent of s and m. Under Assumption 1 in Section 4, with $c_0 = 1$ [which is less general than RE$(s, s, 1)$, cf. Lemma 4.1(i)], a bound of the form (7.6) for the case $p = 2$ is established by Candes and Tao [7].

Bounds on the rate of convergence of the Lasso selector are quite similar to those obtained in Theorem 7.1. They are given by the following result.

THEOREM 7.2. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Let all the diagonal elements of the matrix $X^{\mathrm{T}}X/n$ be equal to* 1, *and let $\mathcal{M}(\beta^*) \le s$, where $1 \le s \le M$, $n \ge 1$, $M \ge 2$. Let Assumption* RE$(s, 3)$ *be satisfied. Consider the Lasso estimator $\widehat{\beta}_L$ defined by* (7.2) *with*

$$r = A\sigma \sqrt{\frac{\log M}{n}}$$

and $A > 2\sqrt{2}$. *Then, with probability at least* $1 - M^{1-A^2/8}$, *we have*

$$|\widehat{\beta}_L - \beta^*|_1 \le \frac{16A}{\kappa^2(s,3)}\sigma s\sqrt{\frac{\log M}{n}}, \tag{7.7}$$

$$|X(\widehat{\beta}_L - \beta^*)|_2^2 \le \frac{16A^2}{\kappa^2(s,3)}\sigma^2 s \log M, \tag{7.8}$$

$$\mathcal{M}(\widehat{\beta}_L) \le \frac{64\phi_{\max}}{\kappa^2(s,3)} s. \tag{7.9}$$

If Assumption RE$(s, m, 3)$ *is satisfied, then, with the same probability as above, simultaneously for all* $1 < p \le 2$, *we have*

$$|\widehat{\beta}_L - \beta^*|_p^p \le 16\left\{1 + 3\sqrt{\frac{s}{m}}\right\}^{2(p-1)} s\left(\frac{A\sigma}{\kappa^2(s,m,3)}\sqrt{\frac{\log M}{n}}\right)^p. \tag{7.10}$$

Inequalities of the form similar to (7.7) and (7.8) can be deduced from the results of [3] under more restrictive conditions on the Gram matrix (the mutual coherence assumption, cf. Assumption 5 of Section 4).

Assumptions RE$(s, 1)$ and RE$(s, 3)$, respectively, can be dropped in Theorems 7.1 and 7.2 if we assume $\beta^* \in \Lambda_{s,\gamma,c_0}$ with $c_0 = 1$ or $c_0 = 3$ as appropriate. Then, (7.4) and (7.5) or, respectively, (7.7) and (7.8) hold with $\kappa = \gamma$. This is analogous to Corollary 6.2. Similarly, (7.6) and (7.10) hold with $\kappa = \gamma$ if $\beta^* \in \Lambda_{s,\gamma,m,c_0}$ with $c_0 = 1$ or $c_0 = 3$ as appropriate.

Observe that, combining Theorems 7.1 and 7.2, we can immediately get bounds for the differences between Lasso and Dantzig selector $|\widehat{\beta}_L - \widehat{\beta}_D|_p^p$ and $|X(\widehat{\beta}_L - \widehat{\beta}_D)|_2^2$. Such bounds have the same form as those of Theorems 7.1 and 7.2, up to numerical constants. Another way of estimating these differences follows directly from the proof of Theorem 7.1. It suffices to observe that the only property of β^* used in that proof is the fact that β^* satisfies the Dantzig constraint on the event of given probability, which is also true for the Lasso solution $\widehat{\beta}_L$. So, we can replace β^* by $\widehat{\beta}_L$ and s by $\mathcal{M}(\widehat{\beta}_L)$ everywhere in Theorem 7.1. Generalizing a bit more, we easily derive the following fact.

THEOREM 7.3. *The result of Theorem* 7.1 *remains valid if we replace* $|\widehat{\beta}_D - \beta^*|_p^p$ *by* $\sup\{|\widehat{\beta}_D - \beta|_p^p : \beta \in \Lambda, \mathcal{M}(\beta) \le s\}$ *for* $1 \le p \le 2$ *and* $|X(\widehat{\beta}_D - \beta^*)|_2^2$ *by* $\sup\{|X(\widehat{\beta}_D - \beta)|_2^2 : \beta \in \Lambda, \mathcal{M}(\beta) \le s\}$, *respectively. Here*, Λ *is the set of all vectors satisfying the Dantzig constraint.*

REMARKS.

1. Theorems 7.1 and 7.2 only give nonasymptotic upper bounds on the loss, with some probability and under some conditions. The probability depends on M and the conditions depend on n and M. Recall that Assumptions RE(s, c_0) and RE(s, m, c_0) are imposed on the $n \times M$ matrix X. To deduce asymptotic conver-

gence (as $n \to \infty$ and/or as $M \to \infty$) from Theorems 7.1 and 7.2, we would need some very strong additional properties, such as simultaneous validity of Assumption RE(s, c_0) or RE(s, m, c_0) (with one and the same constant κ) for infinitely many n and M.

2. Note that neither Assumption RE(s, c_0) or RE(s, m, c_0) implies identifiability of β^* in the linear model (7.1). However, the vector β^* appearing in the statements of Theorems 7.1 and 7.2 is uniquely defined, because we additionally suppose that $\mathcal{M}(\beta^*) \le s$ and $c_0 \ge 1$. Indeed, if there exists a β' such that $X\beta' = X\beta^*$, and $\mathcal{M}(\beta') \le s$, then, in view of assumption RE(s, c_0) with $c_0 \ge 1$, we necessarily have $\beta^* = \beta'$ [cf. discussion following the definition of RE(s, c_0)]. On the other hand, Theorem 7.3 applies to certain values of β that do not come from the model (7.1) at all.

3. For the smallest value of A (which is $A = 2\sqrt{2}$) the constants in the bound of Theorem 7.2 for the Lasso are larger than the corresponding numerical constants for the Dantzig selector given in Theorem 7.1, again, for the smallest admissible value $A = \sqrt{2}$. On the contrary, the Dantzig selector has certain defects as compared to Lasso when the model is nonparametric, as discussed in Section 6. In particular, to obtain sparsity oracle inequalities for the Dantzig selector, we need some restrictions on f, for example, the weak sparsity property. On the other hand, the sparsity oracle inequality (6.1) for the Lasso is valid with no restriction on f.

4. The proofs of Theorems 7.1 and 7.2 differ mainly in the value of the tuning constant, which is $c_0 = 1$ in Theorem 7.1 and $c_0 = 3$ in Theorem 7.2. Note that, since the Lasso solution satisfies the Dantzig constraint, we could have obtained a result similar to Theorem 7.2, but with less accurate numerical constants, by simply conducting the proof of Theorem 7.1 with $c_0 = 3$. However, we act differently, and we deduce (B.30) directly from (B.1) and not from (B.25). This is done only for the sake of improving the constants. In fact, using (B.25) with $c_0 = 3$ would yield (B.30) with the doubled constant on the right-hand side.

5. For the Dantzig selector in the linear regression model and under Assumptions 1 or 2, some further improvement of constants in the ℓ_p bounds for the coefficients can be achieved by applying the general version of Lemma 4.1 with the projector P_{01} inside. We do not pursue this issue here.

6. All of our results are stated with probabilities at least $1 - M^{1-A^2/2}$ or $1 - M^{1-A^2/8}$. These are reasonable (but not the most accurate) lower bounds on the probabilities $\mathbb{P}(\mathcal{B})$ and $\mathbb{P}(\mathcal{A})$, respectively. We have chosen them for readability. Inspection of (B.4) shows that they can be refined to $1 - 2M\Phi(A\sqrt{\log M})$ and $1 - 2M\Phi(A\sqrt{\log M}/2)$, respectively, where $\Phi(\cdot)$ is the standard normal c.d.f.

APPENDIX A

PROOF OF LEMMA 4.1. Consider a partition J_0^c into subsets of size m, with the last subset of size $\le m$: $J_0^c = \bigcup_{k=1}^K J_k$, where $K \ge 1$, $|J_k| = m$ for

$k = 1, \dots, K-1$ and $|J_K| \le m$, such that J_k is the set of indices corresponding to m largest in absolute value coordinates of $\boldsymbol{\delta}$ outside $\bigcup_{j=1}^{k-1} J_j$ (for $k < K$) and J_K is the remaining subset. We have

$$\begin{aligned}
|P_{01} X \boldsymbol{\delta}|_2 &\ge |P_{01} X \boldsymbol{\delta}_{J_{01}}|_2 - \left| \sum_{k=2}^{K} P_{01} X \boldsymbol{\delta}_{J_k} \right|_2 \\
&= |X \boldsymbol{\delta}_{J_{01}}|_2 - \left| \sum_{k=2}^{K} P_{01} X \boldsymbol{\delta}_{J_k} \right|_2 \qquad \text{(A.1)} \\
&\ge |X \boldsymbol{\delta}_{J_{01}}|_2 - \sum_{k=2}^{K} |P_{01} X \boldsymbol{\delta}_{J_k}|_2.
\end{aligned}$$

We will prove first part (ii) of the lemma. Since for $k \ge 1$ the vector $\boldsymbol{\delta}_{J_k}$ has only m nonzero components, we obtain

$$\frac{1}{\sqrt{n}} |P_{01} X \boldsymbol{\delta}_{J_k}|_2 \le \frac{1}{\sqrt{n}} |X \boldsymbol{\delta}_{J_k}|_2 \le \sqrt{\phi_{\max}(m)} |\boldsymbol{\delta}_{J_k}|_2. \tag{A.2}$$

Next, as in [7], we observe that $|\boldsymbol{\delta}_{J_{k+1}}|_2 \le |\boldsymbol{\delta}_{J_k}|_1 / \sqrt{m}$, $k = 1, \dots, K-1$. Therefore,

$$\sum_{k=2}^{K} |\boldsymbol{\delta}_{J_k}|_2 \le \frac{|\boldsymbol{\delta}_{J_0^c}|_1}{\sqrt{m}} \le \frac{c_0 |\boldsymbol{\delta}_{J_0}|_1}{\sqrt{m}} \le c_0 \sqrt{\frac{s}{m}} |\boldsymbol{\delta}_{J_0}|_2 \le c_0 \sqrt{\frac{s}{m}} |\boldsymbol{\delta}_{J_{01}}|_2, \tag{A.3}$$

where we used (4.1). From (A.1)–(A.3), we find

$$\begin{aligned}
\frac{1}{\sqrt{n}} |X \boldsymbol{\delta}|_2 &\ge \frac{1}{\sqrt{n}} |X \boldsymbol{\delta}_{J_{01}}|_2 - c_0 \sqrt{\phi_{\max}(m)} \sqrt{\frac{s}{m}} |\boldsymbol{\delta}_{J_{01}}|_2 \\
&\ge \left(\sqrt{\phi_{\min}(s+m)} - c_0 \sqrt{\phi_{\max}(m)} \sqrt{\frac{s}{m}} \right) |\boldsymbol{\delta}_{J_{01}}|_2,
\end{aligned}$$

which proves part (ii) of the lemma.

The proof of part (i) is analogous. The only difference is that we replace, in the above argument, m by s, and instead of (A.2), we use the bound (cf. [7])

$$\frac{1}{\sqrt{n}} |P_{01} X \boldsymbol{\delta}_{J_k}|_2 \le \frac{\theta_{s,2s}}{\sqrt{\phi_{\min}(2s)}} |\boldsymbol{\delta}_{J_k}|_2. \qquad \square$$

APPENDIX B: TWO LEMMAS AND THE PROOFS OF THE RESULTS

LEMMA B.1. *Fix $M \ge 2$ and $n \ge 1$. Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$, and let $\widehat{f}_L$ be the Lasso estimator defined by* (2.2) *with*

$$r = A\sigma \sqrt{\frac{\log M}{n}}$$

for some $A > 2\sqrt{2}$. *Then, with probability at least* $1 - M^{1-A^2/8}$, *we have, simultaneously for all* $\beta \in \mathbb{R}^M$,

$$
\begin{aligned}
&\|\widehat{f}_L - f\|_n^2 + r\sum_{j=1}^{M}\|f_j\|_n|\widehat{\beta}_{j,L} - \beta_j| \\
&\qquad \le \|f_\beta - f\|_n^2 + 4r\sum_{j\in J(\beta)}\|f_j\|_n|\widehat{\beta}_{j,L} - \beta_j| \qquad \text{(B.1)}\\
&\qquad \le \|f_\beta - f\|_n^2 + 4r\sqrt{\mathcal{M}(\beta)}\sqrt{\sum_{j\in J(\beta)}\|f_j\|_n^2|\widehat{\beta}_{j,L} - \beta_j|^2},
\end{aligned}
$$

and

$$
\left|\frac{1}{n}X^{\mathrm{T}}(\mathbf{f} - X\widehat{\beta}_L)\right|_\infty \le 3rf_{\max}/2. \qquad \text{(B.2)}
$$

Furthermore, with the same probability,

$$
\mathcal{M}(\widehat{\beta}_L) \le 4\phi_{\max} f_{\min}^{-2}(\|\widehat{f}_L - f\|_n^2/r^2), \qquad \text{(B.3)}
$$

where $\phi_{\max}$ *denotes the maximal eigenvalue of the matrix* $X^{\mathrm{T}}X/n$.

PROOF OF LEMMA B.1. The result (B.1) is essentially Lemma 1 from [5]. For completeness, we give its proof. Set $r_{n,j} = r\|f_j\|_n$. By definition,

$$
\widehat{S}(\widehat{\beta}_L) + 2\sum_{j=1}^{M} r_{n,j}|\widehat{\beta}_{j,L}| \le \widehat{S}(\beta) + 2\sum_{j=1}^{M} r_{n,j}|\beta_j|
$$

for all $\beta \in \mathbb{R}^M$, which is equivalent to

$$
\begin{aligned}
&\|\widehat{f}_L - f\|_n^2 + 2\sum_{j=1}^{M} r_{n,j}|\widehat{\beta}_{j,L}| \\
&\qquad \le \|f_\beta - f\|_n^2 + 2\sum_{j=1}^{M} r_{n,j}|\beta_j| + \frac{2}{n}\sum_{i=1}^{n} W_i(\widehat{f}_L - f_\beta)(Z_i).
\end{aligned}
$$

Define the random variables $V_j = n^{-1}\sum_{i=1}^{n} f_j(Z_i)W_i$, $1 \le j \le M$, and the event

$$
\mathcal{A} = \bigcap_{j=1}^{M}\{2|V_j| \le r_{n,j}\}.
$$

Using an elementary bound on the tails of Gaussian distribution, we find that the probability of the complementary event $\mathcal{A}^c$ satisfies

$$
\begin{aligned}
\mathbb{P}\{\mathcal{A}^c\} &\le \sum_{j=1}^{M}\mathbb{P}\{\sqrt{n}|V_j| > \sqrt{n}r_{n,j}/2\} \le M\mathbb{P}\{|\eta| \ge r\sqrt{n}/(2\sigma)\} \\
&\le M\exp\left(-\frac{nr^2}{8\sigma^2}\right) = M\exp\left(-\frac{A^2\log M}{8}\right) = M^{1-A^2/8}, \qquad \text{(B.4)}
\end{aligned}
$$

where $\eta \sim \mathcal{N}(0, 1)$. On the event $\mathcal{A}$ we have

$$\|\widehat{f}_L - f\|_n^2 \le \|f_\beta - f\|_n^2 + \sum_{j=1}^M r_{n,j}|\widehat{\beta}_{j,L} - \beta_j| + \sum_{j=1}^M 2r_{n,j}|\beta_j| - \sum_{j=1}^M 2r_{n,j}|\widehat{\beta}_{j,L}|.$$

Adding the term $\sum_{j=1}^M r_{n,j}|\widehat{\beta}_{j,L} - \beta_j|$ to both sides of this inequality yields, on $\mathcal{A}$,

$$\begin{aligned}\|\widehat{f}_L - f\|_n^2 &+ \sum_{j=1}^M r_{n,j}|\widehat{\beta}_{j,L} - \beta_j| \\ &\le \|f_\beta - f\|_n^2 + 2\sum_{j=1}^M r_{n,j}(|\widehat{\beta}_{j,L} - \beta_j| + |\beta_j| - |\widehat{\beta}_{j,L}|).\end{aligned}$$

Now, $|\widehat{\beta}_{j,L} - \beta_j| + |\beta_j| - |\widehat{\beta}_{j,L}| = 0$ for $j \notin J(\beta)$, so that, on $\mathcal{A}$, we get (B.1).

To prove (B.2) it suffices to note that, on $\mathcal{A}$, we have

$$\left|\frac{1}{n}D^{-1/2}X^{\mathrm{T}}W\right|_\infty \le r/2. \tag{B.5}$$

Now, $\mathbf{y} = \mathbf{f} + \mathbf{w}$, and (B.2) follows from (2.3) and (B.5).

We finally prove (B.3). The necessary and sufficient condition for $\widehat{\beta}_L$ to be the Lasso solution can be written in the form

$$\begin{aligned}&\frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}}(y - X\widehat{\beta}_L) = r\|f_j\|_n \operatorname{sign}(\widehat{\beta}_{j,L}) \qquad \text{if } \widehat{\beta}_{j,L} \ne 0,\\ &\left|\frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}}(y - X\widehat{\beta}_L)\right| \le r\|f_j\|_n \qquad \text{if } \widehat{\beta}_{j,L} = 0,\end{aligned} \tag{B.6}$$

where $\mathbf{x}_{(j)}$ denotes the jth column of X, $j = 1, \ldots, M$. Next, (B.5) yields that, on $\mathcal{A}$, we have

$$\left|\frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}}W\right| \le r\|f_j\|_n/2, \qquad j = 1, \ldots, M. \tag{B.7}$$

Combining (B.6) and (B.7), we get

$$\left|\frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}}(\mathbf{f} - X\widehat{\beta}_L)\right| \ge r\|f_j\|_n/2 \qquad \text{if } \widehat{\beta}_{j,L} \ne 0. \tag{B.8}$$

Therefore,

$$\begin{aligned}\frac{1}{n^2}(\mathbf{f} - X\widehat{\beta}_L)^{\mathrm{T}}XX^{\mathrm{T}}(\mathbf{f} - X\widehat{\beta}_L) &= \frac{1}{n^2}\sum_{j=1}^M \big(\mathbf{x}_{(j)}^{\mathrm{T}}(\mathbf{f} - X\widehat{\beta}_L)\big)^2 \\ &\ge \frac{1}{n^2}\sum_{j:\widehat{\beta}_{j,L}\ne 0} \big(\mathbf{x}_{(j)}^{\mathrm{T}}(\mathbf{f} - X\widehat{\beta}_L)\big)^2 \\ &= \mathcal{M}(\widehat{\beta}_L)r^2\|f_j\|_n^2/4 \ge f_{\min}^2\mathcal{M}(\widehat{\beta}_L)r^2/4.\end{aligned}$$

Since the matrices $X^{\mathrm{T}}X/n$ and XX^{T}/n have the same maximal eigenvalues,

$$\frac{1}{n^2}(\mathbf{f}-X\widehat{\beta}_L)^{\mathrm{T}}XX^{\mathrm{T}}(\mathbf{f}-X\widehat{\beta}_L)\le\frac{\phi_{\max}}{n}|\mathbf{f}-X\widehat{\beta}_L|_2^2=\phi_{\max}\|f-\widehat{f}_L\|_n^2,$$

and we deduce (B.3) from the last two displays. □

COROLLARY B.2. *Let the assumptions of Lemma* B.1 *be satisfied and* $\|f_j\|_n=1$, $j=1,\dots,M$. *Consider the linear regression model* $\mathbf{y}=X\beta+\mathbf{w}$. *Then, with probability at least* $1-M^{1-A^2/8}$, *we have*

$$|\boldsymbol{\delta}_{J_0^c}|_1\le 3|\boldsymbol{\delta}_{J_0}|_1,$$

where $J_0=J(\beta)$ *is the set of nonzero coefficients of* β *and* $\boldsymbol{\delta}=\widehat{\beta}_L-\beta$.

PROOF. Use the first inequality in (B.1) and the fact that $f=f_\beta$ for the linear regression model. □

LEMMA B.3. *Let* $\beta\in\mathbb{R}^M$ *satisfy the Dantzig constraint*

$$\left|\frac{1}{n}D^{-1/2}X^{\mathrm{T}}(y-X\beta)\right|_\infty\le r$$

and set $\boldsymbol{\delta}=\widehat{\beta}_D-\beta$, $J_0=J(\beta)$. *Then,*

$$|\boldsymbol{\delta}_{J_0^c}|_1\le|\boldsymbol{\delta}_{J_0}|_1. \tag{B.9}$$

Further, let the assumptions of Lemma B.1 *be satisfied with* $A>\sqrt{2}$. *Then, with probability of at least* $1-M^{1-A^2/2}$, *we have*

$$\left|\frac{1}{n}X^{\mathrm{T}}(\mathbf{f}-X\widehat{\beta}_D)\right|_\infty\le 2rf_{\max}. \tag{B.10}$$

PROOF OF LEMMA B.3. Inequality (B.9) follows immediately from the definition of Dantzig selector (cf. [7]). To prove (B.10), consider the event

$$\mathcal{B}=\left\{\left|\frac{1}{n}D^{-1/2}X^{\mathrm{T}}W\right|_\infty\le r\right\}=\bigcap_{j=1}^{M}\{|V_j|\le r_{n,j}\}.$$

Analogously to (B.4), $\mathbb{P}\{\mathcal{B}^c\}\le M^{1-A^2/2}$. On the other hand, $\mathbf{y}=\mathbf{f}+\mathbf{w}$, and, using the definition of Dantzig selector, it is easy to see that (B.10) is satisfied on $\mathcal{B}$. □

PROOF OF THEOREM 5.1. Set $\boldsymbol{\delta}=\widehat{\beta}_L-\widehat{\beta}_D$. We have

$$\frac{1}{n}|\mathbf{f}-X\widehat{\beta}_L|_2^2=\frac{1}{n}|\mathbf{f}-X\widehat{\beta}_D|_2^2-\frac{2}{n}\boldsymbol{\delta}^{\mathrm{T}}X^{\mathrm{T}}(\mathbf{f}-X\widehat{\beta}_D)+\frac{1}{n}|X\boldsymbol{\delta}|_2^2.$$

This and (B.10) yield

$$
\begin{aligned}
\|\widehat{f}_D - f\|_n^2 &\le \|\widehat{f}_L - f\|_n^2 + 2|\boldsymbol{\delta}|_1 \left|\frac{1}{n}X^{\mathrm{T}}(\mathbf{f} - X\widehat{\beta}_D)\right|_\infty - \frac{1}{n}|X\boldsymbol{\delta}|_2^2 \\
&\le \|\widehat{f}_L - f\|_n^2 + 4f_{\max}r|\boldsymbol{\delta}|_1 - \frac{1}{n}|X\boldsymbol{\delta}|_2^2,
\end{aligned}
\tag{B.11}
$$

where the last inequality holds with probability at least $1 - M^{1-A^2/2}$. Since the Lasso solution $\widehat{\beta}_L$ satisfies the Dantzig constraint, we can apply Lemma B.3 with $\beta = \widehat{\beta}_L$, which yields

$$
|\boldsymbol{\delta}_{J_0^c}|_1 \le |\boldsymbol{\delta}_{J_0}|_1 \tag{B.12}
$$

with $J_0 = J(\widehat{\beta}_L)$. By Assumption RE($s$, 1), we get

$$
\frac{1}{\sqrt{n}}|X\boldsymbol{\delta}|_2 \ge \kappa|\boldsymbol{\delta}_{J_0}|_2, \tag{B.13}
$$

where $\kappa = \kappa(s, 1)$. Using (B.12) and (B.13), we obtain

$$
|\boldsymbol{\delta}|_1 \le 2|\boldsymbol{\delta}_{J_0}|_1 \le 2\mathcal{M}^{1/2}(\widehat{\beta}_L)|\boldsymbol{\delta}_{J_0}|_2 \le \frac{2\mathcal{M}^{1/2}(\widehat{\beta}_L)}{\kappa\sqrt{n}}|X\boldsymbol{\delta}|_2. \tag{B.14}
$$

Finally, from (B.11) and (B.14), we get that, with probability at least $1 - M^{1-A^2/2}$,

$$
\begin{aligned}
\|\widehat{f}_D - f\|_n^2 &\le \|\widehat{f}_L - f\|_n^2 + \frac{8f_{\max}r\mathcal{M}^{1/2}(\widehat{\beta}_L)}{\kappa\sqrt{n}}|X\boldsymbol{\delta}|_2 - \frac{1}{n}|X\boldsymbol{\delta}|_2^2 \\
&\le \|\widehat{f}_L - f\|_n^2 + \frac{16f_{\max}^2 r^2\mathcal{M}(\widehat{\beta}_L)}{\kappa^2},
\end{aligned}
\tag{B.15}
$$

where the RHS follows (B.2), (B.10) and another application of (B.14). This proves one side of the inequality.

To show the other side of the bound on the difference, we act as in (B.11), up to the inversion of roles of $\widehat{\beta}_L$ and $\widehat{\beta}_D$, and we use (B.2). This yields that, with probability at least $1 - M^{1-A^2/8}$,

$$
\begin{aligned}
\|\widehat{f}_L - f\|_n^2 &\le \|\widehat{f}_D - f\|_n^2 + 2|\boldsymbol{\delta}|_1 \left|\frac{1}{n}X^{\mathrm{T}}(\mathbf{f} - X\widehat{\beta}_L)\right|_\infty - \frac{1}{n}|X\boldsymbol{\delta}|_2^2 \\
&\le \|\widehat{f}_D - f\|_n^2 + 3f_{\max}r|\boldsymbol{\delta}|_1 - \frac{1}{n}|X\boldsymbol{\delta}|_2^2.
\end{aligned}
\tag{B.16}
$$

This is analogous to (B.11). Now, paralleling the proof leading to (B.15), we obtain

$$
\|\widehat{f}_L - f\|_n^2 \le \|\widehat{f}_D - f\|_n^2 + \frac{9f_{\max}^2 r^2\mathcal{M}(\widehat{\beta}_L)}{\kappa^2}. \tag{B.17}
$$

The theorem now follows from (B.15) and (B.17). $\square$

PROOF OF THEOREM 5.2. Set, again, $\boldsymbol{\delta} = \widehat{\beta}_L - \widehat{\beta}_D$. We apply (B.1) with $\beta = \widehat{\beta}_D$, which yields that, with probability at least $1 - M^{1-A^2/8}$,

$$|\boldsymbol{\delta}|_1 \le 4|\boldsymbol{\delta}_{J_0}|_1 + \|\widehat{f}_D - f\|_n^2/r, \tag{B.18}$$

where, now, $J_0 = J(\widehat{\beta}_D)$. Consider the following two cases: (i) $\|\widehat{f}_D - f\|_n^2 > 2r|\boldsymbol{\delta}_{J_0}|_1$ and (ii) $\|\widehat{f}_D - f\|_n^2 \le 2r|\boldsymbol{\delta}_{J_0}|_1$. In case (i), inequality (B.16) with $f_{\max} = 1$ immediately implies

$$\|\widehat{f}_L - f\|_n^2 \le 10\|\widehat{f}_D - f\|_n^2,$$

and the theorem follows. In case (ii), we get, from (B.18), that

$$|\boldsymbol{\delta}|_1 \le 6|\boldsymbol{\delta}_{J_0}|_1$$

and thus $|\boldsymbol{\delta}_{J_0^c}|_1 \le 5|\boldsymbol{\delta}_{J_0}|_1$. We can therefore apply Assumption RE$(s, 5)$, which yields, similarly to (B.14),

$$|\boldsymbol{\delta}|_1 \le 6\mathcal{M}^{1/2}(\widehat{\beta}_D)|\boldsymbol{\delta}_{J_0}|_2 \le \frac{6\mathcal{M}^{1/2}(\widehat{\beta}_D)}{\kappa\sqrt{n}}|X\boldsymbol{\delta}|_2, \tag{B.19}$$

where $\kappa = \kappa(s, 5)$. Plugging (B.19) into (B.16) we finally get that, in case (ii),

$$\begin{aligned}\|\widehat{f}_L - f\|_n^2 &\le \|\widehat{f}_D - f\|_n^2 + \frac{18r\mathcal{M}^{1/2}(\widehat{\beta}_D)}{\kappa\sqrt{n}}|X\boldsymbol{\delta}|_2 - \frac{1}{n}|X\boldsymbol{\delta}|_2^2 \\ &\le \|\widehat{f}_D - f\|_n^2 + \frac{81r^2\mathcal{M}(\widehat{\beta}_D)}{\kappa^2}.\end{aligned} \tag{B.20}$$

□

PROOF OF THEOREM 6.1. Fix an arbitrary $\beta \in \mathbb{R}^M$ with $\mathcal{M}(\beta) \le s$. Set $\boldsymbol{\delta} = D^{1/2}(\widehat{\beta}_L - \beta)$, $J_0 = J(\beta)$. On the event $\mathcal{A}$, we get, from the first line in (B.1), that

$$\begin{aligned}\|\widehat{f}_L - f\|_n^2 + r|\boldsymbol{\delta}|_1 &\le \|f_\beta - f\|_n^2 + 4r\sum_{j \in J_0}\|f_j\|_n|\widehat{\beta}_{j,L} - \beta_j| \\ &= \|f_\beta - f\|_n^2 + 4r|\boldsymbol{\delta}_{J_0}|_1,\end{aligned} \tag{B.21}$$

and from the second line in (B.1) that

$$\|\widehat{f}_L - f\|_n^2 \le \|f_\beta - f\|_n^2 + 4r\sqrt{\mathcal{M}(\beta)}|\boldsymbol{\delta}_{J_0}|_2. \tag{B.22}$$

Consider, separately, the cases where

$$4r|\boldsymbol{\delta}_{J_0}|_1 \le \varepsilon\|f_\beta - f\|_n^2 \tag{B.23}$$

and

$$\varepsilon\|f_\beta - f\|_n^2 < 4r|\boldsymbol{\delta}_{J_0}|_1. \tag{B.24}$$

In case (B.23), the result of the theorem trivially follows from (B.21). So, we will only consider the case (B.24). All of the subsequent inequalities are valid on the event $\mathcal{A} \cap \mathcal{A}_1$, where $\mathcal{A}_1$ is defined by (B.24). On this event, we get, from (B.21), that

$$|\boldsymbol{\delta}|_1 \le 4(1 + 1/\varepsilon)|\boldsymbol{\delta}_{J_0}|_1,$$

which implies $|\boldsymbol{\delta}_{J_0^c}|_1 \le (3 + 4/\varepsilon)|\boldsymbol{\delta}_{J_0}|_1$. We now use Assumption RE$(s, 3 + 4/\varepsilon)$. This yields

$$\begin{aligned}\kappa^2|\boldsymbol{\delta}_{J_0}|_2^2 &\le \frac{1}{n}|X\boldsymbol{\delta}|_2^2 = \frac{1}{n}(\widehat{\beta}_K - \beta)^{\mathrm{T}} D^{1/2} X^{\mathrm{T}} X D^{1/2}(\widehat{\beta}_L - \beta)\\ &\le \frac{f_{\max}^2}{n}(\widehat{\beta}_L - \beta)^{\mathrm{T}} X^{\mathrm{T}} X(\widehat{\beta}_L - \beta) = f_{\max}^2 \|\widehat{f}_L - f_\beta\|_n^2,\end{aligned}$$

where $\kappa = \kappa(s, 3 + 4/\varepsilon)$. Combining this with (B.22), we find

$$\begin{aligned}\|\widehat{f}_L - f\|_n^2 &\le \|f_\beta - f\|_n^2 + 4rf_{\max}\kappa^{-1}\sqrt{\mathcal{M}(\beta)}\|\widehat{f}_L - f_\beta\|_n\\ &\le \|f_\beta - f\|_n^2 + 4rf_{\max}\kappa^{-1}\sqrt{\mathcal{M}(\beta)}(\|\widehat{f}_L - f\|_n + \|f_\beta - f\|_n).\end{aligned}$$

This inequality is of the same form as (A.4) in [4]. A standard decoupling argument as in [4], using inequality $2xy \le x^2/b + by^2$ with $b > 1$, $x = r\kappa^{-1}\sqrt{\mathcal{M}(\beta)}$ and y being either $\|\widehat{f}_L - f\|_n$ or $\|f_\beta - f\|_n$, yields that

$$\|\widehat{f}_L - f\|_n^2 \le \frac{b+1}{b-1}\|f_\beta - f\|_n^2 + \frac{8b^2 f_{\max}^2}{(b-1)\kappa^2} r^2 \mathcal{M}(\beta) \qquad \forall b > 1.$$

Taking $b = 1 + 2/\varepsilon$ in the last display finishes the proof of the theorem. □

PROOF OF PROPOSITION 6.3. Due to the weak sparsity assumption, there exists $\bar{\beta} \in \mathbb{R}^M$ with $\mathcal{M}(\bar{\beta}) \le s$ such that $\|f_{\bar{\beta}} - f\|_n^2 \le C_0 f_{\max}^2 r^2 \kappa^{-2} \mathcal{M}(\bar{\beta})$, where $\kappa = \kappa(s, 3 + 4/\varepsilon)$ is the same as in Theorem 6.1. Using this together with Theorem 6.1 and (B.3), we obtain that, with probability at least $1 - M^{1-A^2/8}$,

$$\mathcal{M}(\widehat{\beta}_L) \le C_1(\varepsilon)\mathcal{M}(\bar{\beta}) \le C_1(\varepsilon)s.$$

This and Theorem 5.1 imply

$$\|\widehat{f}_D - f\|_n^2 \le \|\widehat{f}_L - f\|_n^2 + \frac{16C_1(\varepsilon) f_{\max}^2 A^2\sigma^2}{\kappa_0^2}\left(\frac{s\log M}{n}\right),$$

where $\kappa_0 = \kappa(\max(C_1(\varepsilon), 1)s, 3 + 4/\varepsilon)$. Once Again, applying Theorem 6.1, we get the result. □

PROOF OF THEOREM 7.1. Set $\boldsymbol{\delta} = \widehat{\beta}_D - \beta^*$ and $J_0 = J(\beta^*)$. Using Lemma B.3 with $\beta = \beta^*$, we get that, on the event $\mathcal{B}$ (i.e., with probability at least

$1 - M^{1-A^2/2}$), the following are true: (i) $\frac{1}{n}|X^{\mathrm{T}}X\boldsymbol{\delta}|_\infty \le 2r$, and (ii) inequality (4.1) holds with $c_0 = 1$. Therefore, on $\mathcal{B}$ we have

$$\begin{aligned}
\frac{1}{n}|X\boldsymbol{\delta}|_2^2 &= \frac{1}{n}\boldsymbol{\delta}^{\mathrm{T}}X^{\mathrm{T}}X\boldsymbol{\delta} \\
&\le \frac{1}{n}|X^{\mathrm{T}}X\boldsymbol{\delta}|_\infty|\boldsymbol{\delta}|_1 \\
&\le 2r(|\boldsymbol{\delta}_{J_0}|_1 + |\boldsymbol{\delta}_{J_0^c}|_1) \\
&\le 2(1+c_0)r|\boldsymbol{\delta}_{J_0}|_1 \\
&\le 2(1+c_0)r\sqrt{s}|\boldsymbol{\delta}_{J_0}|_2 = 4r\sqrt{s}|\boldsymbol{\delta}_{J_0}|_2
\end{aligned} \tag{B.25}$$

since $c_0 = 1$. From Assumption RE$(s, 1)$, we get that

$$\frac{1}{n}|X\boldsymbol{\delta}|_2^2 \ge \kappa^2|\boldsymbol{\delta}_{J_0}|_2^2,$$

where $\kappa = \kappa(s, 1)$. This and (B.25) yield that, on $\mathcal{B}$,

$$\frac{1}{n}|X\boldsymbol{\delta}|_2^2 \le 16r^2 s/\kappa^2, \qquad |\boldsymbol{\delta}_{J_0}|_2 \le 4r\sqrt{s}/\kappa^2. \tag{B.26}$$

The first inequality in (B.26) implies (7.5). Next, (7.4) is straightforward in view of the second inequality in (B.26) and of the relations (with $c_0 = 1$)

$$|\boldsymbol{\delta}|_1 = |\boldsymbol{\delta}_{J_0}|_1 + |\boldsymbol{\delta}_{J_0^c}|_1 \le (1+c_0)|\boldsymbol{\delta}_{J_0}|_1 \le (1+c_0)\sqrt{s}|\boldsymbol{\delta}_{J_0}|_2 \tag{B.27}$$

that hold on $\mathcal{B}$. It remains to prove (7.6). It is easy to see that the kth largest in absolute value element of $\boldsymbol{\delta}_{J_0^c}$ satisfies $|\boldsymbol{\delta}_{J_0^c}|_{(k)} \le |\boldsymbol{\delta}_{J_0^c}|_1/k$. Thus,

$$|\boldsymbol{\delta}_{J_{01}^c}|_2^2 \le |\boldsymbol{\delta}_{J_0^c}|_1^2 \sum_{k\ge m+1}\frac{1}{k^2} \le \frac{1}{m}|\boldsymbol{\delta}_{J_0^c}|_1^2,$$

and, since (4.1) holds on $\mathcal{B}$ (with $c_0 = 1$), we find

$$|\boldsymbol{\delta}_{J_{01}^c}|_2 \le \frac{c_0|\boldsymbol{\delta}_{J_0}|_1}{\sqrt{m}} \le c_0|\boldsymbol{\delta}_{J_0}|_2\sqrt{\frac{s}{m}} \le c_0|\boldsymbol{\delta}_{J_{01}}|_2\sqrt{\frac{s}{m}}.$$

Therefore, on $\mathcal{B}$,

$$|\boldsymbol{\delta}|_2 \le \left(1 + c_0\sqrt{\frac{s}{m}}\right)|\boldsymbol{\delta}_{J_{01}}|_2. \tag{B.28}$$

On the other hand, it follows from (B.25) that

$$\frac{1}{n}|X\boldsymbol{\delta}|_2^2 \le 4r\sqrt{s}|\boldsymbol{\delta}_{J_{01}}|_2.$$

Combining this inequality with Assumption RE$(s, m, 1)$, we obtain that, on $\mathcal{B}$,

$$|\boldsymbol{\delta}_{J_{01}}|_2 \le 4r\sqrt{s}/\kappa^2.$$

Recalling that $c_0 = 1$ and applying the last inequality together with (B.28), we get

$$|\delta|_2^2 \le 16\left(1 + c_0\sqrt{\frac{s}{m}}\right)^2 (r\sqrt{s}/\kappa^2)^2. \tag{B.29}$$

It remains to note that (7.6) is a direct consequence of (7.4) and (B.29). This follows from the fact that inequalities $\sum_{j=1}^M a_j \le b_1$ and $\sum_{j=1}^M a_j^2 \le b_2$ with $a_j \ge 0$ imply

$$\begin{aligned}\sum_{j=1}^M a_j^p &= \sum_{j=1}^M a_j^{2-p} a_j^{2p-2} \le \left(\sum_{j=1}^M a_j\right)^{2-p} \left(\sum_{j=1}^M a_j^2\right)^{p-1} \\ &\le b_1^{2-p} b_2^{p-1} \qquad \forall 1 < p \le 2.\end{aligned}$$ □

PROOF OF THEOREM 7.2. Set $\boldsymbol{\delta} = \widehat{\beta}_L - \beta^*$ and $J_0 = J(\beta^*)$. Using (B.1), where we put $\beta = \beta^*$, $r_{n,j} \equiv r$ and $\|f_\beta - f\|_n = 0$, we get that, on the event $\mathcal{A}$,

$$\frac{1}{n}|X\boldsymbol{\delta}|_2^2 \le 4r\sqrt{s}|\boldsymbol{\delta}_{J_0}|_2 \tag{B.30}$$

and (4.1) holds with $c_0 = 3$ on the same event. Thus, by Assumption RE(s, 3) and the last inequality, we obtain that, on $\mathcal{A}$,

$$\frac{1}{n}|X\boldsymbol{\delta}|_2^2 \le 16r^2 s/\kappa^2, \qquad |\boldsymbol{\delta}_{J_0}|_2 \le 4r\sqrt{s}/\kappa^2, \tag{B.31}$$

where $\kappa = \kappa(s, 3)$. The first inequality here coincides with (7.8). Next, (7.9) follows immediately from (B.3) and (7.8). To show (7.7), it suffices to note that on the event $\mathcal{A}$ the relations (B.27) hold with $c_0 = 3$, to apply the second inequality in (B.31) and to use (B.4).

Finally, the proof of (7.10) follows exactly the same lines as that of (7.6). The only difference is that one should set $c_0 = 3$ in (B.28) and (B.29), as well as in the display preceding (B.28). □

REFERENCES

[1] BICKEL, P. J. (2007). Discussion of "The Dantzig selector: Statistical estimation when p is much larger than n," by E. Candes and T. Tao. *Ann. Statist.* **35** 2352–2357. MR2382645

[2] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2004). Aggregation for regression learning. Preprint LPMA, Univ. Paris 6–Paris 7, n° 948. Available at arXiv:math.ST/0410214 and at https://hal.ccsd.cnrs.fr/ccsd-00003205.

[3] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2006). Aggregation and sparsity via ℓ_1 penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006)* (G. Lugosi and H. U. Simon, eds.). *Lecture Notes in Artificial Intelligence* **4005** 379–391. Springer, Berlin. MR2280619

[4] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. MR2351101

[5] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Statist.* **1** 169–194. MR2312149

[6] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Sparse density estimation with ℓ_1 penalties. In *Proceedings of 20th Annual Conference on Learning Theory (COLT 2007)* (N. H. Bshouty and C. Gentile, eds.). *Lecture Notes in Artificial Intelligence* **4539** 530–543. Springer, Berlin. MR2397610

[7] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[8] DONOHO, D. L., ELAD, M. and TEMLYAKOV, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory* **52** 6–18. MR2237332

[9] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–451. MR2060166

[10] FRIEDMAN, J., HASTIE, T., HÖFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1** 302–332. MR2415737

[11] FU, W. and KNIGHT, K. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

[12] GREENSHTEIN, E. and RITOV, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli* **10** 971–988. MR2108039

[13] JUDITSKY, A. and NEMIROVSKI, A. (2000). Functional aggregation for nonparametric estimation. *Ann. Statist.* **28** 681–712. MR1792783

[14] KOLTCHINSKII, V. (2006). Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincaré Probab. Statist.* To appear.

[15] KOLTCHINSKII, V. (2007). Dantzig selector and sparsity oracle inequalities. Unpublished manuscript.

[16] MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The Group Lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B* **70** 53–71.

[17] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[18] MEINSHAUSEN, N. and YU, B. (2006). Lasso type recovery of sparse representations for high dimensional data. *Ann. Statist.* To appear.

[19] NEMIROVSKI, A. (2000). Topics in nonparametric statistics. In *Ecole d'Eté de Probabilités de Saint-Flour XXVIII—1998. Lecture Notes in Math.* **1738**. Springer, New York. MR1775640

[20] OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A (2000a). On the Lasso and its dual. *J. Comput. Graph. Statist.* **9** 319–337. MR1822089

[21] OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A (2000b). A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **20** 389–404.

[22] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[23] TSYBAKOV, A. B. (2006). Discussion of "Regularization in Statistics," by P. Bickel and B. Li. *TEST* **15** 303–310. MR2273731

[24] TURLACH, B. A. (2005). On algorithms for solving least squares problems under an L1 penalty or an L1 constraint. In *2004 Proceedings of the American Statistical Association, Statistical Computing Section [CD-ROM]* 2572–2577. Amer. Statist. Assoc., Alexandria, VA.

[25] VAN DE GEER, S. A. (2008). High dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 614–645. MR2396809

[26] ZHANG, C.-H. and HUANG, J. (2008). Model-selection consistency of the Lasso in high-dimensional regression. *Ann. Statist.* **36** 1567–1594. MR2435448

[27] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

P. J. BICKEL
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA AT BERKELEY
CALIFORNIA
USA
E-MAIL: bickel@stat.berkeley.edu

Y. RITOV
DEPARTMENT OF STATISTICS
FACULTY OF SOCIAL SCIENCES
THE HEBREW UNIVERSITY
JERUSALEM 91904
ISRAEL
E-MAIL: yaacov.ritov@gmail.com

A. B. TSYBAKOV
LABORATOIRE DE STATISTIQUE
CREST
3, AVENUE PIERRE LAROUSSE,
92240 MALAKOFF
AND
LPMA (UMR CNRS 1599)
UNIVERSITÈ PARIS VI
4, PLACE JUSSIEU,
75252 PARIS, CEDEX 05
FRANCE
E-MAIL: alexandre.tsybakov@upmc.fr

Chapter 8
Miscellaneous

Ya'acov Ritov

8.1 Introduction to Four Papers by Peter Bickel

8.1.1 General Introduction

We introduce here four paper coauthored by P. J. Bickel. These papers have very little in common. Two of them can be considered mainly as papers dealing with concepts, while the two others are mainly tedious hard technical work that aims in developing complicated probabilistic results, which can be applied to the asymptotic theory of estimators.

8.1.2 Minimax Estimation of the Mean of a Normal Distribution When the Parameter Space Is Restricted

The paper "Minimax Estimation of the Mean of a Normal Distribution when the Parameter Space is Restricted", Bickel (1981), discusses mainly a very simple problem, which is almost a textbook problem. Suppose that $X \sim N(\theta, 1)$, θ should be estimated with a quadratic loss function. So far, this is the most trivial example of an estimation problem, where X is the minimax decision. However, when it is known a priori that $|\theta| \leq m$, for some $m \in (0, \infty)$, the problem is not anymore trivial. In fact, prior to 1981, the answer was known only m small enough (slightly larger than 1). The minimax decision then is the Bayes decision with respect to prior which puts all its mass on the two end points of the interval.

Y. Ritov (✉)
Department of Statistics, Hebrew University of Jerusalem, Jerusalem, Israel
e-mail: yaacov@mscc.huji.ac.il

J. Fan et al. (eds.), *Selected Works of Peter J. Bickel*, Selected Works in Probability and Statistics 13, DOI 10.1007/978-1-4614-5544-8_8,

Now, the following claims are relatively simple.

1. We consider the "game" between the statistician and Nature, in which Nature selects $\theta \in [-m,m]$ according to some π. The statistician observes $X \sim N(\theta,1)$, selects a real $d(X)$, and then he pays Nature $\big(d(X)-\theta\big)^2$. This game has a saddle points (π,d). Clearly, given π, d should be δ_π, the Bayes with respect to π. The existence of π follows from a general argument involving continuity, convexity, and compactness.
2. Since π is a maxmin strategy for Nature. Its support is included in $A(d) = \{s : |s| \le m, R_s(\delta_\pi) = \max_t R_t(\delta_\pi)\}$, where $R_s(d)$ is the risk of the decision d at s.
3. $A(d)$ is a finite set. This follows since for any d, $R_\cdot(d)$ is analytic in s, and hence there cannot be a dense set in which $R_\cdot(\delta_\pi)$ achieves its maximum. On the other hand the support of π cannot be too sparse, because then it would be likely that θ is the support point closest to X.

Peter makes these observations precise, and then characterizes the asymptotic behavior of π. He shows that if $\pi = \pi_m$, then $m\pi_m(ms)$ converges weakly to the distribution on $(-1,1)$ with density $g(s) = \cos^2(\pi s/2)$. Moreover the asymptotic risk is $1-\pi^2/m^2+o(m^{-2})$.

This result is generalized to the multivariate case, where the prior is restricted to a ball. It would be generalized then further by Melkman and Ritov (1987) to a general real distribution, but notably by Donoho et al. (1990) for a very general asymptotic result.

8.1.3 What Is a Linear Process?

The paper "What is a linear process?" (Bickel and Bühlmann 1996) shows that modeling an empirical phenomena may be tricky. Testing for abstract notion like stationarity is, essentially, impossible. Looking on a long time series and say, 'Clearly, it is not stationary', is not necessarily possible.

A linear process, or a moving average, is defined to be the stationary process

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad t = \ldots,-1,0,1,\ldots,$$

where ε_i are i.i.d., with mean 0 and finite variance, and $\psi_1,\psi_2,\ldots$ are given constants with a finite sum of squares. Since the authors consider an infinite moving average, their model includes the causal autoregressive process. The authors define some natural topologies over these process, and consider the closure of this set. The closure includes all objects that naturally should be there like MA process and all mean 0 Gaussian process. But it includes a surprising type of processes. To describe this set, we consider independent processes: $\xi_{\ldots,i,j}$, where, for each i, $\xi_{\cdot,i,1},\xi_{\cdot,i,2},\ldots$ are i.i.d. copies of a stationary process, and then consider the set of all processes that can be described as $X_\cdot = \sum_{i=1}^{\infty}\sum_{j=1}^{N_i}\xi_{\cdot,i,j}$, where $N_1,N_2,\ldots$ are independent (non-identical) Poisson random variables.

This latter type of limit is what makes the paper exciting. In Fig. 1, ten realization of a MA process with a finite window size. In the different graphs a realization of sample size which is ten times the window size is given. The realization of the same process look impressively different. Recall that different realizations behave like far away pieces of the same process.

How all this related to testing? Fact 1.3 of the paper is very clear: "In testing the hypothesis H_O about MA representation against any fixed one-point alternative H_A about a nonlinear, stationary process, there is no test with asymptotic significance level $\alpha < 0.36$ having limiting power 1 as the sample size tends to infinity."

8.1.4 Sums of Functions of Nearest Neighbor Distances, Moment Bounds, Limit Theorems and a Goodness of Fit Test

It is simple to see that if $X_1,\dots,X_n$ are i.i.d. from a the uniform $U(0,1)$ distribution, the spacing between the observations behave like a sample from exponential distribution. More generally, if they are a sample from a distribution with a smooth density $f(\cdot)$, and $R_i = \min|X_j - X_i|$, $j \neq i$, then R_i is a the minimum of two independent exponential random variables with mean $1/f(X_i)$, that is $2R_i$ is asymptotically like an exponential random variable with mean $1/f(X_i)$.

This was relatively simple, because, we could use the probability transformation to assume WLOG that the random variables are uniform. And then it is well known that the time of the events of a Poisson process divided by the time of the n-th event are distributed like the order statistics of a sample from the $U(0,1)$ distribution. But, how much can this results extended to the general case of $m > 1$ dimension?

The somewhat surprising result, given by Bickel and Breiman (1983), is that this is true. If, similarly to the above, $R_i = \min\|X_j - X_i\|$, $j \neq i$, and $V(r)$ is the volume of the m-dimensional ball of radius r, then $W_i = \exp\big(-nf(X_i)V(R_i)\big)$, $i = 1,\dots,n$ behave like a sample from the uniform distribution. In fact, the paper shows that if $\hat{H}$ is the empirical cumulative distribution function of $W_1,\dots,W_n$, then $Z_n(t) = \sqrt{n}\big(\hat{H}(t) - E\hat{H}(t)\big)$ converges weakly to a zero mean Gaussian process whose covariance does not depend on the f. It makes sense that $\hat{H}$ is asymptotically normal, since although the W_i's are not independent, there is enough mixing and far away points are almost independent. The actual proof is hard and a long complicated paper was needed.

8.1.5 Convergence Criteria for Multiparameter Stochastic Processes and Some Applications

In the fourth paper of this section, Bickel and Wichura (1971) generalized a univariate result of Billingsley (1968), in which weak convergence of $D(0,1)$ processes. They took the ideas from Billingsley, and prove results which may be not the stronger, and may lead to not to the most elegant prove, but they are typically cheap and based mostly on moments on the fluctuation of the functions. The difficulty is as above in moving from the well ordered world of the real line, to the general Euclidean space, where boundaries are not finite.

To get the general feeling of the result we quote Nielsen and Linton (1995), which gives a simplified result:

Lemma 8.1. *Let $X(t)$ be a stochastic process with $t = (t_1, \ldots, t_d) \in [0,1]^d$. For any $t \in [0,1]^d$ and $v \in [0,1]$, let $t_{j,v} = (t_1, \ldots, t_{j-1}, v, t_{j+1}, \ldots, t_d)$. If for $C > 0$: $X(t) \xrightarrow{p} 0$ for all $t \in [0,1]^d$, and $E\big(X(t) - X(t_{j,u})\big)^2 < C(t_j - u)^2$ for all $t \in [0,1]^d$ and $j = 1, \ldots, d$. Then $\sup|X(t)| \xrightarrow{p} 0$.*

Nielsen and Linton then go and prove the uniform convergence of the hazard rate estimate in a nonparametric setup with time dependent hazard rate, with multivariate regressors.

Acknowledgements This work was partially supported by an ISF grant.

References

Bickel PJ (1981) Minimax estimation of the mean of a normal distribution when the parameter space is restricted. Ann Stat 9:1301–1309
Bickel PJ, Breiman L (1983) Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. Ann Probab 11:185–214
Bickel PJ, Bühlmann P (1996) What is a linear process? Proc Natl Acad Sci 93:12128–12131
Bickel PJ, Wichura M (1971) Convergence criteria for multiparameter stochastic processes and some applications. Ann Math Stat 42:1656–1669
Billingsley P (1968) Convergence of probability measures. Wiley, New York
Donoho DL, Liu RC, MacGibbon B (1990) Minimax estimation over hyperrectangles with implications. Ann Stat 18:1416–1437
Melkman AA, Ritov Y (1987) Minimax estimation of the mean of a general distribution when the parameter space is restricted. Ann Stat 15:432–442
Nielsen JP, Linton OB (1995) Kernel estimation in a nonparametric marker dependent Hazard model. Ann Stat 23:1735–1748

The Annals of Mathematical Statistics
1971, Vol. 42, No. 5, 1656–1670

CONVERGENCE CRITERIA FOR MULTIPARAMETER STOCHASTIC PROCESSES AND SOME APPLICATIONS[1]

BY P. J. BICKEL AND M. J. WICHURA
University of California, Berkeley and University of Chicago

Chentsov–Billingsley type fluctuation inequalities for stochastic processes whose time parameter ranges over the q-dimensional unit cube are derived and used to establish weak convergence results for such processes.

1. Introduction. In his excellent recent book (1968), Billingsley has given several fluctuation inequalities for sums of random variables (Theorems 12.1, 12.2, 12.5, 12.6) leading to convergence criteria for sequences of stochastic processes $(X_n(t))_{t\in[0,1]}$ whose sample paths are right-continuous and have left-limits everywhere. These criteria, which may be viewed as generalizations of results of Kolmogorov and Chentsov (1956), have been applied by Billingsley to provide simple proofs of various classical results in the theory of weak convergence of one-parameter stochastic processes.

There has recently been considerable interest in questions of weak convergence of similar stochastic processes $(X_n(t))$, where t ranges over the unit cube in q-dimensional space. Situations in which such convergence arises include:

(i) Convergence of the normalized empirical cumulative distribution function for samples from a continuous distribution concentrating on the unit cube in R^q (Dudley (1966), Le Cam (1957)).

(ii) Convergence of the analogue of the partial sum process for two and higher dimensional "time" (Kuelbs (1968), Wichura (1969)).

(iii) Convergence of the normalized, randomly-stopped empirical cumulative for samples from a q-dimensional continuous distribution on the unit q-cube (Pyke (1968), Wichura (1968)).

(iv) Convergence of the normalized empirical cumulative for samples (drawn without replacement) from a finite population (Bickel (1969), Rosén (1967)).

In this paper we prove multidimensional analogues of Theorems 12.5 and 15.6 of Billingsley (1968) and apply them in the situations cited above. The fluctuation inequalities may be found in Section 2 in a format similar to that given in Billingsley (1968) pages 87–102, the convergence criteria in Section 3, and the applications in Section 4.

Other methods work, frequently more elegantly, in all of the above examples. However, as in the one-dimensional case, in situations where moments are "cheap" and the dependence structure formidable we feel that this approach will prove important. In particular we hope to show in a subsequent paper how these criteria

Received November 4, 1970.

[1] This paper was prepared with the partial support of ONR Contract N00014–67–A–011–0004, and also by NSF Grant NSF–6P16071 and ONR Contract N00014–67–A–0285–0009.

may be successfully applied to the problem of convergence of the normalized, randomly-stopped, empirical cumulative distribution of the normalized sample spacings from a uniform distribution on [0, 1]. The question of whether this sequence of processes converges weakly was posed by Pyke (1965).

2. Fluctuation inequalities. Let q be a positive integer, and let $T_1, \cdots, T_q$ be subsets of [0, 1], each of which contains 0 and 1, and is either a finite set or [0, 1] itself. Put $T = T_1 \times \cdots \times T_q$. Let $X = (X(t))_{t \in T}$ be a stochastic process whose state space is some linear space E (typically R^1) endowed with a norm, say $|\cdot|$; we assume that the sample paths of X are smooth enough to permit each of the supremal quantities defined below to be computed by running the time indices involved through countable dense subsets. For simplicity, we assume that X vanishes along the lower boundary, $\bigcup_{1 \leqq p \leqq q} T_1 \times \cdots \times T_{p-1} \times \{0\} \times T_{p+1} \times \cdots \times T_q$, of T. For each p and each $t \in T_p$ define $X_t^{(p)}: T_1 \times \cdots \times T_{p-1} \times T_{p+1} \times \cdots \times T_q \to E$ by

$$X_t^{(p)}(t_1, \cdots, t_{p-1}, t_{p+1}, \cdots, t_q) = X(t_1, \cdots, t_{p-1}, t, t_{p+1}, \cdots, t_q),$$

and for each $s \leqq t \leqq u$ in T_p, set

$$m_p(s, t, u) \equiv m_p(s, t, u)(X) = \min\left(\|X_t^{(p)} - X_s^{(p)}\|, \|X_u^{(p)} - X_t^{(p)}\|\right),$$

where $\|\cdot\|$ is the usual supremum norm. The quantities of primary concern to us here are the random variables

$$M_p'' \equiv M_p''(X) = \sup\{m_p(s, t, u): s \leqq t \leqq u \in T_p\}$$

$(1 \leqq p \leqq q)$ and

$$M'' \equiv M''(X) = \max_p M_p''.$$

For $p = 1$ and T finite, the modulus M'' is that of Billingsley (1968) (cf. (12.62)), which is very useful in studying the weak convergence of $D([0, 1])$-valued processes. Our goal in this section is to establish bounds on the tail probabilities of the M_p'''s, and thus also on those of M''.

In passing, we note that bounds on M'' give rise to bounds on the random variable

$$M \equiv \sup\{|X(t)|: t \in T\}$$

via the inequality (compare Billingsley (12.4))

$$M \leqq \textstyle\sum_{1 \leqq p \leqq q} M_p'' + |X(u)| \leqq qM'' + |X(u)|, \tag{1}$$

where $u = (1, \cdots, 1)$. To establish this inequality take any $t = (t_1, \cdots, t_q) \in T$, and set $u_p = (1, \cdots, 1, t_{p+1}, \cdots, t_q)(0 \leqq p \leqq q)$, so that $u_0 = t$ and $u_q = u$. The assumption that X vanishes along the lower boundary of T then yields

$$\begin{aligned}|X(u_{p-1})| &\leqq \min\{|X(u_{p-1})|, |X(u_p) - X(u_{p-1})|\} + |X(u_p)| \\ &\leqq M_p'' + |X(u_p)|\end{aligned}$$

for $1 \leqq p \leqq q$; together, these inequalities imply that $|X(t)|$ is majorized by the middle term of (1).

To describe the hypotheses under which we will derive the desired bounds, we will make use of the following notation and terminology. *A* **block** B in T is a subset of T of the form $(s, t] = \prod_p (s_p, t_p]$ with s and t in T; the pth-**face** of $B = (s, t]$ is $\prod_{\rho \neq p}(s_\rho, t_\rho]$. Disjoint blocks B and C are p-**neighbors** if they abut and have the same pth face; they are **neighbors** if they are p-neighbors for some p (for example, when $q = 3$, the blocks $(s, t] \times (a, b] \times (c, d]$ and $(t, u] \times (a, b] \times (c, d]$ are 1-neighbors ($s \leqq t \leqq u$ in T_1)). For each block $B = (s, t]$, let

$$X(B) = \sum_{\varepsilon_1 = 0,1} \cdots \sum_{\varepsilon_q = 0,1} (-1)^{q - \Sigma_p \varepsilon_p} X(s_1 + \varepsilon_1(t_1 - s_1), \cdots, s_q + \varepsilon_q(t_q - s_q))$$

be the **increment** of X around B; $X(\cdot)$ is a (random) finitely additive function on blocks. For each pair of neighboring blocks B, C, put

$$m(B, C) = \min\{|X(B)|, |X(C)|\},$$

$m(B, C)$ is small iff at least one of the increments $X(B)$ and $X(C)$ is small.

Now let $\beta > 1$ and $\gamma > 0$, and let μ be a finite nonnegative measure on T. Again for simplicity, we assume that μ assigns measure zero to the lower boundary of T. Say that (X, μ) **satisfies condition** (β, γ), and write $(X, \mu) \in \mathscr{C}(\beta, \gamma)$, if

$$P\{m(B, C) \geqq \lambda\} \leqq \lambda^{-\gamma}(\mu(B \cup C))^\beta \tag{2}$$

for all $\lambda > 0$ and every pair of neighboring blocks B and C in T. From Chebychev's inequality, one sees that (2) is implied by its moment version, namely $E(m(B, C))^\gamma \leqq (\mu(B \cup C))^\beta$, as well as by the frequently employed moment condition

$$E(|X(B)|^{\gamma_1}|X(C)|^{\gamma_2}) \leqq (\mu(B))^{\beta_1}(\mu(C))^{\beta_2}, \tag{3}$$

where γ_1, γ_2, β_1, and β_2 satisfy $\gamma_1 + \gamma_2 = \gamma$ and $\beta_1 + \beta_2 = \beta$. When $q = 1$ and T is finite, condition (2) is essentially (12.11) of Billingsley (with β here equal to 2α there, and γ here equal to 2γ there).

Define constants $K_q(\beta, \gamma)$ and $L_q(\beta, \gamma)$ inductively as follows: Put $\delta = 1/(1+\gamma)$, $\rho = 2^{-(\beta-1)\delta}$, $K(\beta, \gamma) = 2^\gamma(1-\rho)^{-1/\delta}$, $K_1(\beta, \gamma) = L_1(\beta, \gamma) = 2^\beta K(\beta, \gamma)$, and for $r \geqq 2$, $K_r(\beta, \gamma) = K_1(\beta, \gamma)([L_{r-1}(\beta, \gamma)(r-1)^\gamma]^\delta + 1)^{1/\delta}$, $L_r(\beta, \gamma) = rK_r(\beta, \gamma)$. Here is the main result, which for $q = 1$ is a variant both of Billingsley's Theorem 12.5 and Chentsov's (1956) Theorem 1.

THEOREM 1. *If* $(X, \mu) \in \mathscr{C}(\beta, \gamma)$, *then*

$$P\{M_p''(X) \geqq \lambda\} \leqq K_q(\beta, \gamma)\lambda^{-\gamma}(\mu(T))^\beta \qquad (1 \leqq p \leqq q) \tag{4}$$

$$P\{M''(X) \geqq \lambda\} \leqq L_q(\beta, \gamma)\lambda^{-\gamma}(\mu(T))^\beta \tag{5}$$

for all positive λ.

A few remarks should be made at this point. When $T = [0, 1]^q$ and μ is continuous, the factor 2^β may be dropped from the definition of $K_1(\beta, \gamma)$, thus giving smaller universal constants. For T finite and $q = 1$, Theorem 1 reduces to

Theorem 12.5 of Billingsley, except that Billingsley gives a different value, namely $2^{(2+\gamma-\beta)}K_1(\beta, \gamma)$, for the universal constant. The K's are quite large; for example, when $q = 1$ and, as in many applications, $\beta = 2$ and $\gamma = 4$, $K_1(\beta, \gamma)$ is approximately 1,750,000. Finally, the assumption that the process X and the measure μ vanish along the lower boundary of T can be removed, provided condition (β, γ) is strengthened so as to restrain the behavior of X over the lower boundary; what is needed is simply that (X', μ') satisfy condition (β, γ), where (slightly abusing our convention concerning time domains) $T' = T_1' \times \cdots \times T_q'$, $T_p' = \{-1\} \cup T_p$, and X' (resp. μ') equals X (resp. μ) over T and zero over $T' \sim T$ (note that $M_T''(X) \leqq M_T''(X')$).

PROOF OF THEOREM 1. The proof will be carried out in several steps, as follows: (i) $q = 1$, $T = [0, 1]$, $\mu =$ Lebesgue measure, (ii) $q = 1$, $T = [0, 1]$, μ atomless, (iii) $q = 1$, T finite, (iv) $q = 1$, $T = [0, 1]$, μ general, and (v) $q \geqq 2$.

Step 1. Here condition (β, γ) reads

$$P[\min\{|X(t)-X(s)|, |X(u)-X(t)|\} \geqq \lambda] \leqq \lambda^{-\gamma}(u-s)^{\beta} \tag{6}$$

for all $\lambda > 0$ and all $0 \leqq s \leqq t \leqq u \leqq 1$; we shall show that (6) implies

$$P\{M'' \geqq \lambda\} \leqq K(\beta, \gamma)\lambda^{-\gamma}$$

for all $\lambda > 0$.

Take any positive numbers θ_i, $i \geqq 0$, set

$$s_{i,n} = (n-1)2^{-i}, u_{i,n} = n2^{-i}, t_{i,n} = (s_{i,n}+u_{i,n})/2,$$

and define events

$$F_{i,n} = \{\min(|X(t_{i,n})-X(s_{i,n})|, |X(u_{i,n})-X(t_{i,n})|) < \lambda\theta_i\}$$
$$F_i = \textstyle\bigcap_{1 \leqq n \leqq 2^i} F_{i,n}$$
$$F = \textstyle\bigcup_{0 \leqq i < \infty} F_i.$$

If $F_{i,n}$ occurs, then one has a "favorable" comparison of the two increments involved, in the sense that at least one of them is "small."

On the one hand, the probability that all comparisons are favorable is high, i.e.

$$P(F^c) \leqq \textstyle\sum_i \sum_n P(F_{i,n}^c) \leqq \sum_i 2^i(\lambda\theta_i)^{-\gamma}2^{-i\beta} = \lambda^{-\gamma}\sum_{0 \leqq i < \infty} 2^{-\alpha i}\theta_i^{-\gamma}, \tag{7}$$

where $\alpha = \beta - 1 > 0$. On the other hand, whenever all comparisons are favorable, M'' is small, i.e.

$$F \subset \{M'' \leqq 2(\textstyle\sum_{0 \leqq i < \infty} \theta_i)\lambda\}. \tag{8}$$

To see this, let $S_i = \{n2^{-i}; 0 \leqq n \leqq 2^i\}$, let $\omega \in F$, and, referring to the definition of the F_i's, construct ω-dependent order-preserving maps $\psi_i: S_{i+1} \to S_i$ such that

$$|X(\psi_i(s))(\omega) - X(s)(\omega)| < \lambda\theta_i$$

for all $s \in S_{i+1}$ and all i. Piece the ψ_i's together to produce an (ω-dependent) order-preserving map ψ from $S = \bigcup_i S_i$ to $\{0, 1\}$ such that

$$|X(\psi(s))(\omega) - X(s)(\omega)| < \lambda \textstyle\sum_{0 \leqq i < \infty} \theta_i$$

for all $s \in S$. By the monotonicity of ψ, one must have $\psi(s) = \psi(t)$ or $\psi(t) = \psi(u)$ for any three points $s \leqq t \leqq u$ in S. Our assumption about the smoothness of the sample paths of X now implies that (8) holds.

From (7) and (8), one sees that M'' is likely to be small, i.e.

$$P\{M'' \geqq \lambda\} \leqq \lambda^{-\gamma} \inf_\xi f(\xi),$$

where $\xi = (\xi_i)_{i \geqq 0}$ ranges over all probability measures on $\{0, 1, 2, \cdots\}$ and where

$$f(\xi) = 2^\gamma \textstyle\sum_{i \geqq 0} 2^{-\alpha i} \xi_i^{-\gamma}.$$

Elementary calculations show that f achieves its minimum at that ξ for which $\xi_i = \rho^i(1-\rho)$ (all i) and has there the value $K(\beta, \gamma)$.

Step 2. *Proof for* $q = 1$, $T = [0, 1]$, μ *having continuous distribution function* F. For F both continuous and strictly increasing, a transformation of the time scale making use of the well-defined inverse function of F reduces the present case to that treated in Step 1, and yields

$$P\{M'' \geqq \lambda\} \leqq K(\beta, \gamma)\lambda^{-\gamma}(F(1))^\beta \tag{9}$$

($F(0) = 0$ by assumption). For F merely continuous, first note that (9) holds with F replaced by $F + \varepsilon I$ (I = identity function), and then pass to the limit as $\varepsilon \downarrow 0$.

Step 3. *Proof for* $q = 1$, T *finite*. Let $0 = t_0 < t_1 < \cdots < t_m = 1$ be the points of T. Let $Y = (Y(u))_{0 \leqq u \leqq 1}$ be the process, defined on the same probability space as X, having right continuous sample paths constant over the intervals separating the t_i's and satisfying $Y(t_i) = X(t_i)$ for $0 \leqq i \leqq m$, i.e.

$$Y(u) = \textstyle\sum_{0 \leqq i < m} X(t_i) I_{[t_i, t_{i+1})}(u) + X(t_m) I_{\{t_m\}}(u).$$

One has $M_T''(X) = M''_{[0,1]}(Y)$. Now look at $m(s, t, u)(Y)$. This quantity is zero unless

$$0 \leqq t_{i-1} \leqq s < t_i \leqq t < t_k \leqq u < t_{k+1}$$

for some $0 < i < k \leqq m$, in which case the hypotheses on X yield

$$\begin{aligned} P\{m(s,t,u)(Y) \geqq \lambda\} &\leqq \lambda^{-\gamma} \left(\textstyle\sum_{i \leqq j \leqq k} \mu(\{t_j\})\right)^\beta \\ &\leqq \lambda^{-\gamma} \left[\textstyle\sum_{i < j \leqq k} (\mu(\{t_j\}) + \mu(\{t_{j-1}\}))\right]^\beta \\ &\leqq \lambda^{-\gamma} [F(u) - F(s)]^\beta, \end{aligned}$$

where F is that continuous distribution function, satisfying $F(0) = 0$, which is linear over $[t_{j-1}, t_j]$ with

$$F(t_j) - F(t_{j-1}) = \mu(\{t_j\}) + \mu(\{t_{j-1}\})$$

for $1 \leqq j \leqq m$. Since $F(1) \leqq 2\mu(T)$, it follows from Step 2 that

$$P\{M''(X) \geqq \lambda\} \leqq \lambda^{-\gamma} K_1(\beta, \gamma)(\mu(T))^\beta.$$

Step 4. *Proof for* $q = 1$, $T = [0, 1]$ (μ *arbitrary*). Let $0 = t_0 < t_1 < \cdots < t_m = 1$ be points in T, put $U = \{t_0, \cdots, t_m\}$, $Y = (X(u))_{u \in U}$, and let ν be the measure on U such that

$$\begin{aligned} \nu(\{t_j\}) &= \mu((t_{j-1}, t_j]), && \text{if } j \geqq 1 \\ &= 0, && \text{if } j = 0. \end{aligned}$$

Since (X, μ) satisfies condition (β, γ), so does (Y, ν). Apply Step 3 to (Y, ν) and make a suitable passage to the limit to get the desired result.

Step 5. *Proof for* $q \geqq 2$. We now know that Theorem 1 is true when $q = 1$, i.e. when we are dealing with univariate time. The rest of the proof proceeds by induction on q for (4) and (5) simultaneously. Consider (4), with $p = 1$ for convenience. The key observation to be made is that the univariate-time version of Theorem 1 may be applied to the (function space valued) process $(X_t^{(1)})_{t \in T_1}$, once bounds on the increments of this process are found; these bounds will come to us from (1) and the induction hypothesis. More specifically, let $s \leqq t \leqq u$ in T_1, and define processes $Y = X_t^{(1)} - X_s^{(1)}$ and $Z = X_u^{(1)} - X_t^{(1)}$ having $T_2 \times \cdots \times T_q$ as index set. From (1), we have

$$M(Y) \leqq (q-1)M''(Y) + |Y(\mathbf{1})|, M(Z) \leqq (q-1)M''(Z) + |Z(\mathbf{1})|$$

(where $\mathbf{1} = (1, 1, \cdots, 1)$), so that

$$m_1(s,t,u)(X) = \min\{M(Y), M(Z)\} \leqq (q-1)[\max\{M''(Y), M''(Z)\}] + \min\{|Y(\mathbf{1})|, |Z(\mathbf{1})|\}.$$

Using the fact that the increment of Y around a block B in $T_2 \times \cdots \times T_q$ is the increment of X around the block $(s, t] \times B$ in T, one gets from the induction hypothesis that

$$P\{M''(Y) \geqq \lambda\} \leqq \lambda^{-\gamma} L_{q-1}(\beta, \gamma)(F(t) - F(s))^\beta,$$

where F is the distribution function of the marginal of μ on T_1. Similarly,

$$P\{M''(Z) \geqq \lambda\} \leqq \lambda^{-\gamma} L_{q-1}(\beta, \gamma)(F(u) - F(t))^\beta$$

while from the original hypothesis on X,

$$P[\min\{|Y(\mathbf{1})|, |Z(\mathbf{1})|\} \geqq \lambda] \leqq \lambda^{-\gamma}(F(u) - F(s))^\beta.$$

It follows easily from this, the estimate

$$P\{U + V \geqq \lambda\} \leqq P\{U \geqq \lambda\xi_1\} + P\{V \geqq \lambda\xi_2\}$$

(valid for any random variables U, V and positive numbers ξ_1, ξ_2 such that $\xi_1 + \xi_2 = 1$), and the relation

$$\inf\{C_1/\xi_1^\gamma + C_2/\xi_2^\gamma : \xi_1 + \xi_2 = 1\} = (C_1^\delta + C_2^\delta)^{1/\delta}$$

($\delta = 1/(1+\gamma)$) that

$$P\{m_1(s,t,u)(X) \geqq \lambda\} \leqq \lambda^{-\gamma}([(q-1)^\gamma L_{q-1}(\beta, \gamma)]^\delta + 1)^{1/\delta}(F(u) - F(s))^\beta.$$

In other words, the process $X^{(1)}$ meets the hypotheses of the theorem for the case of univariate time, so that (4) holds; of course, (4) implies (5). □

3. Convergence criteria. Let T denote the unit cube $[0, 1]^q$. Call a function $x: T \to R^1$ a step function if x is a linear combination of functions of the form

$$t \to I_{E_1 \times E_2 \times \cdots \times E_q}(t), \qquad \text{where}$$

each E_p is either a left-closed, right-open subinterval of $[0, 1]$, or the singleton $\{1\}$ and where I_E denotes the indicator of the set E. Let D_q be the uniform closure, in the space of all bounded functions from T to R^1, of the vector subspace of simple functions. The functions in D_q may be characterized by their continuity properties, as follows. If $t \in T$ and if, for $1 \leqq p \leqq q$, R_p is one of the relations $<$ and $\geqq$, let $Q_{R_1, \cdots, R_q}(t)$ denote the quadrant

$$\{(s_1, \cdots, s_q) \in T: s_p R_p t_p, 1 \leqq p \leqq q\}.$$

Then (see Neuhaus (1969), or Straf (1970), page 29) $x \in D_q$ iff for each $t \in T$, (a) $x_Q \equiv \lim_{s \to t, s \in Q} x(s)$ exists for each of the 2_q quadrants $Q = Q_{R_1, \cdots, R_q}(t)$, and (b) $x(t) = x_{Q_{\geqq, \ldots, \geqq}}$. In this sense, the functions of $\mathscr{D}_q$ are "continuous from above, with limits from below."

One can introduce a metric topology on D_q which for $q = 1$ coincides with Skorohod's well-known and useful J_1-topology (see Billingsley (1968), for example). For this, let Λ be the group of all transformations $\lambda: T \to T$ of the form $\lambda(t_1, \cdots, t_q) = (\lambda_1(t_1), \cdots, \lambda_q(t_q))$, where each $\lambda_p: [0, 1] \to [0, 1]$ is continuous, strictly increasing, and fixes zero and one. Define the "Skorohod" distance between x and y in D_q to be

$$d(x, y) = \inf\{\min(\|x - y\lambda\|, \|\lambda\|): \lambda \in \Lambda\},$$

where $\|x - y\lambda\| = \sup\{|x(t) - y(\lambda(t))|: t \in T\}$ and $\|\lambda\| = \sup\{|\lambda(t) - t|: t \in T\}$. With respect to the corresponding metric topology (S-topology), D_q is separable and topologically complete, and the Borel σ-algebra $\mathscr{D}_q$ coincides with the σ-algebra generated by the coordinate mappings (Billingsley (1968), Neuhaus (1969), Straf (1969)). Consequently, a stochastic process $(X(t))_{t \in T}$ taking values in D_q is $\mathscr{D}_q$-measurable.

We turn now to a discussion of weak convergence for D_q-valued processes. For simplicity we shall speak only of sequences of processes, but everything we say is true for generalized sequences, i.e. nets. A sequence $(X_n)_{n \geqq 1}$ of D_q-valued processes is said to converge weakly in the S-topology to a D_q-valued process X, written $X_n \to X$, if $Ef(X_n) \to Ef(X)$ for all S-continuous bounded functions $f: D_q \to R$. According to the general theory of weak convergence, $X_n \to X$ is equivalent to $f(X_n) \to f(X)$ (in the sense of weak convergence for real-valued random variables) for all $\mathscr{D}_q$-measurable functions $f: D_q \to R$ which are X-continuous in the S-topology (i.e., continuous almost surely with respect to the distribution of X). If X takes all its values in C_q, the subset of D_q consisting of continuous functions, then one has $f(X_n) \to f(X)$ even for $\mathscr{D}_q$-measurable functions f which are X-continuous with respect to the stronger topology of uniform convergence (see Billingsley (1968), Neuhaus (1969) and Straf (1969)).

A criterion for the weak convergence of D_q-valued processes can be given in terms of the weak convergence of the corresponding finite-dimensional distributions together with a tightness condition. To make this explicit, define π_S: $D_q \to R^S$ by $\pi_S(x) = (x(s))_{s \in S}$, for each finite set $S \subset T$. Let $\mathscr{T}$ be the collection of subsets of T of the form $U_1 \times \cdots \times U_q$, where each U_p contains zero and one and has countable complement. For each D_q-valued process X, put $T_X = \{t \in T, \pi_{\{t\}}$ is continuous with probability one with respect to the law of X on $(D_q, \mathscr{D}_q)\}$; one can show $T_X \in \mathscr{T}$ (Billingsley (1968), Neuhaus (1969), Straf (1969)). Finally, call a partition of T formed by finitely many hyperplanes parallel to the coordinate axes a δ-grid if each element of the partition is a "left-closed, right-open" rectangle of diameter at least δ, and define w_δ': $D_q \to R$ by

$$w_\delta'(x) = \inf_\Delta \max_{G \in \Delta} \sup_{s,t \in G} |x(t) - x(s)|,$$

where the infimum extends over all δ-grids Δ in T. Following the development of Billingsley (1969), it is easy to prove the following fundamental result (confer Straf (1970) page 36):

THEOREM 2. *Let X_n, $n \geqq 1$, be D_q-valued processes. In order that the sequence (X_n) converge weakly, it is necessary and sufficient that*

(i) $(\pi_S(X_n))$ *converges weakly, for all finite subsets S of some member τ of $\mathscr{T}$, and*

(ii) $\operatorname{plim}_\delta \lim_n w_\delta'(X_n) = 0$;

and then $X_n \to X$, where the distribution of the D_q-valued process X is determined by $\pi_S(X_n) \to \pi_S(X)$ for all finite $S \in \tau \cap T_X$. (Condition (ii) means $\lim_{\delta \downarrow 0} \limsup_n P\{w_\delta'(X_n) \geqq \varepsilon\} = 0$ for all $\varepsilon > 0$).

One can deduce (cf Theorem 14.4 and 15.4 of Billingsley (1968)) from this basic result the corollary below, which is sufficient for our purposes. First define w_δ'' : $D_q \to R$ by

$$w_\delta''(x) = \max_p w_\delta''^{(p)}(x)$$

where

$$w_\delta''^{(p)}(x) = \sup \{\min(\|x_t^{(p)} - x_s^{(p)}\|, \|x_u^{(p)} - x_t^{(p)}\|) : s \leqq t \leqq u, u - s \leqq \delta\}$$

$(1 \leqq p \leqq q)$. To motivate this definition, we note that the set-theoretic identity

$$D_q \equiv D(I^q, R) = D_1(I, D_{q-1})$$

is valid via any one of the correspondences $x(\cdot) \leftrightarrow x_\cdot^{(p)}(\cdot)$, provided on the right-hand side D_{q-1} is equipped with the supremum norm. This is easily proved (confer Straf (1970), page 32) by first considering step functions and then their uniform limits. The modulus w_δ'' can thus be viewed as a more or less natural generalization of Billingsley's modulus, of the same name, for $p = 1$. Another consequence of the above identity is that for any D_q-valued process X, $\lim_{t \uparrow 1} X_t^{(p)}$ exists uniformly over $[0, 1]^{q-1}$. This limit will be $X_1^{(p)}$ provided the finite-dimensional distributions of the $X_t^{(p)}$'s converge to those of $X_1^{(p)}$, as will be the case if, say, $X(s_1, \cdots, s_{p-1}, t, s_{p+1}, \cdots, s_q)$ converges to $X(s_1, \cdots, s_{p-1}, 1, s_{p+1}, \cdots, s_q)$ in probability for all choices of the s_j's. We shall say that X is continuous at the upper boundary of T if $\lim_{t\uparrow 1} X_t^{(p)} = X_1^{(p)}$ for each p, with probability one.

COROLLARY. *Let X_n, $n \geqq 1$, and X be D_q-valued processes, and suppose that X is continuous at the upper boundary of T. Then in order that $X_n \to X$, it is necessary and sufficient that*

$$\pi_S(X_n) \to \pi_S(X) \qquad \textit{for all finite subsets } S \textit{ of some member } \tau \textit{ of } \mathscr{T},$$

$$\operatorname{plim}_\delta \lim_n w_\delta''(X_n) = 0. \tag{10}$$

PROOF. Here is the proof of the sufficiency. The proof uses induction on q. For $q = 1$, the corollary is just Theorem 15.4 of Billingsley (1968). Suppose now that the sufficiency part of the corollary is known to hold for $q-1$; we shall show that it holds for q. We have only to verify that condition (ii) of Theorem 2 holds. For each p, define $w_\delta'^{(p)}$ on D_q by

$$w_\delta{}^{(p)}(x) = \inf_{\Delta_p} \max_{G\in\Delta_p} \sup_{s,t\in G} \|x_t^{(p)} - x_s^{(p)}\|,$$

where the infimum here extends over all δ-grids Δ_p in [0, 1]. Clearly,

$$w_\delta' \leqq \textstyle\sum_{1\leqq p\leqq q} w_\delta'^{(p)}.$$

Moreover, a simple but tedious argument (cf Billingsley (1968), Theorems 14.4 and 15.4) shows that

$$w_{\delta/2}'^{(p)}(x) \leqq 2[w_\delta''^{(p)}(x) + L_\delta^{(p)}(x) + R_\delta^{(p)}(x)],$$

where

$$L_\delta^{(p)}(x) = \sup_{0\leqq t<\delta} \|x_t^{(p)} - x_0^{(p)}\| \leqq 2[\|x_\delta^{(p)} - x_0^{(p)}\| + w_\delta''^{(p)}(x)]$$

$$R_\delta^{(p)}(x) = \sup_{\zeta<t\leqq 1} \|x_1^{(p)} - x_t^{(p)}\| \leqq 2[\|x_1^{(p)} - x_\zeta^{(p)}\| + w_\delta''^{(p)}(x)]$$

($\zeta = 1-\delta$).

Thus it suffices to show that the plim $\inf_\delta \lim_n$'s of

$$\|(X_n)_\delta^{(p)} - (X_n)_0^{(p)}\| \quad \text{and} \quad \|(X_n)_1^{(p)} - (X_n)_\zeta^{(p)}\|$$

are zero. As the arguments in both cases are similar, we shall discuss only the first case. Fix p, and set $Z_{n,\delta} = (X_n)_\delta^{(p)} - (X_n)_0^{(p)}$, $Z_\delta = X_\delta^{(p)} - X_0^{(p)}$; Z_δ and the $Z_{n,\delta}$'s are D_{q-1}-valued processes. We will show below that $Z_{n,\delta} \to Z_\delta$ for all but countably many δ's. Since $\|\cdot\| = d(\cdot, 0)$ is an S-continuous function on D_{q-1}, we will then have $\|Z_{n,\delta}\| \to \|Z_\delta\|$ for all but countably many δ's. But the identity $D_q = D_1(I, D_{q-1})$ implies that $\|Z_\delta\| \to 0$ as $\delta \to 0$. All this gives

$$\liminf_\delta \limsup_n P\{\|Z_{n,\delta}\| \geqq \varepsilon\} = 0$$

for all $\varepsilon > 0$, as desired.

It remains to show that $Z_{n,t} \to Z_t$ for all but countably many t. One finds easily that

(a) for any t, the process Z_t is continuous at the upper boundary of $[0, 1]^{q-1}$,

(b) one has $w_\delta''(Z_{n,t}) \leqq 2w_\delta''(X_n)$, where on the left-hand side w_δ'' is the modulus appropriate for D_{q-1}, and, with $\tau = U_1 \times \cdots \times U_q$,

(c) for any $t \in U_1$, one has $\pi_S(Z_{n,t}) \to \pi_S(Z_t)$ for all finite subsets S of $U_2 \times \cdots \times U_q$.

The $q-1$ dimensional version of the corollary now implies that for $t \in U_1$, $Z_{n,t} \to Z_t$. □

For these results to be useful in practice, one needs easily verifiable conditions which imply the somewhat awkward tightness condition (10). This is where the fluctuation inequality of the previous section comes into play. The following theorem extends Theorem 15.6 of Billingsley.

THEOREM 3. *Suppose that each X_n vanishes along the lower boundary of T, and that there exist constants $\beta > 1$, $\gamma > 0$ and a finite nonnegative measure μ on T with continuous marginals such that $(X_n, \mu) \in \mathscr{C}(\beta, \gamma)$ for each n. Then the tightness condition* (10) *is in force.*

PROOF. It is enough to show $\operatorname{plim}_\delta \lim_n w_\delta''^{(p)}(X_n) = 0$ for each p. For this, put

$$w(\sigma, \tau; n) = \sup\{\min(\|(X_n)_t^{(p)} - (X_n)_s^{(p)}\|, \|(X_n)_u^{(p)} - (X_n)_t^{(p)}\|) : \sigma \leqq s \leqq t \leqq u \leqq \tau\}.$$

Since

$$w''^{(p)}_{\frac{1}{2}k}(X_n) \leqq \max_{1 \leqq j \leqq k} w((2j-2)/2k, 2j/2k; n) + \max_{1 \leqq j < k} w((2j-1)/2k, (2j+1)/2k; n)$$

it suffices (cf Billingsley (1968) page 130) to show that

$$P\{w(\sigma, \tau; n) \geqq \varepsilon\} \leqq \varepsilon^{-\gamma} K_q(\beta, \gamma)(\mu_p((\sigma, \tau]))^\beta \tag{11}$$

where μ_p denotes the (continuous) marginal of μ on the pth edge of T. But (11) is an easy consequence of Theorem 1 and the fact that in the definition of $w(\sigma, \tau; n)$, X_n can be replaced by Y_n, where Y_n, defined on $T^* = [0, 1]^{p-1} \times [\sigma, \tau] \times [0, 1]^{q-p}$ so that $(Y_n)_t^{(p)} = (X_n)_t^{(p)} - (X_n)_\sigma^{(p)}$ for $\sigma \leqq t \leqq \tau$, vanishes along the lower boundary of T^* and has the same increments around blocks in T^* as does X_n. □

Actually, Theorem 3 is not flexible enough to apply to some of the simplest processes. The following extension will be useful. For each n, suppose that there exists a subset $T^n = T_1{}^n \times \cdots \times T_q{}^n$ of T such that

(a) $T_p{}^n$ contains 0 and 1 for each n $(1 \leqq p \leqq q)$,

(b) $w_\delta''(X_n)$ may be computed using T^n as the time set (instead of T),

(c) T^n becomes dense in T as n grows large, and

(d) Condition (β, γ) holds for blocks whose corner points lie in T^n.

Then the conclusion to Theorem 3 holds; the proof is essentially the same (with the role of the equally spaced points $j/2k$, $0 \leqq j \leqq 2k$, in the estimate of $w''^{(p)}_{\frac{1}{2}k}(X_n)$ being taken over by almost equally spaced points from $T_p{}^n$). The theorem may be extended further by allowing μ to depend on n and to have discontinuous marginals, while requiring that the new μ_n's converge weakly to a limit μ having continuous marginals (under this condition, (11) holds with a lim $\sup_n$ prefixed to the left-hand side; inspection of the argument on page 130 of Billingsley (1968) shows that this is good enough). Finally we note that an analogue of Theorem 15.7 of Billingsley (1968) can be proved by essentially the same method, and thus that there is no loss of generality in considering only D_q-valued processes from the outset. Specifically one has

THEOREM 4. *Let $\mathscr{S}$ denote the class of finite subsets of T. Let $(\nu_S)_{S\in\mathscr{S}}$ be a consistent family of probabilities on the finite-dimensional spaces $(R^S, \mathscr{R}^S)$, $S \in \mathscr{S}$. Define ν on the algebra $\bigcup_{S\in\mathscr{S}}\pi_S^{-1}(\mathscr{R}^S)$ of subsets of R^T so that $\nu\pi_S^{-1} = \nu_S$ for all S in $\mathscr{S}$. Suppose that*

(i) $\nu\{x \in R^T : x(t) = 0\} = 1$, *if any coordinate of $t \in T$ is* 0,

(ii) $\nu\{x \in R^T : |x(t+h)-x(t)| \geqq \varepsilon\} \to 0$ *for all $\varepsilon > 0$, as h tends to 0 "from above,"*

(iii) $\nu\{x \in R^T : |x(s_1, \cdots, s_{p-1}, t, s_{p+1}, \cdots, s_q)-x(s_1, \cdots, s_{p-1}, 1, s_{p+1}, \cdots, s_q)| \geqq \varepsilon\} \to 0$ *as $t \to 1$, for all choices of p and of the s_j's, $j \neq p$, and for all $\varepsilon > 0$,*

(iv) *for some $\beta > 1$, $\gamma > 0$, and some measure μ on T having continuous marginals,*

$$\nu\{x\in R^T : \min(|x(B)|, |x(C)|) \geqq \lambda\} \leqq \lambda^{-\gamma}(\mu(B \cup C))^\beta$$

for all $\lambda > 0$ and all pairs of neighboring blocks B and C in T.
Then there exists a D_q-valued process whose finite dimensional distributions are the ν_S's.

4. Applications. Our purpose here is to illustrate the use of Theorem 3 in establishing weak convergence results. Accordingly, no fuss will be made about convergence of finite-dimensional distributions, which in most of the examples below is obvious. Some of the results have been deduced before, by a variety of different methods.

(I) *Partial sum processes.* For convenience, we work with 2-dimensional time. The following theorem extends the classic result of Prohorov (for $q = 1$) (cf Prohorov (1956) and Wichura (1969)). For each n, let $X_{i,j}^{(n)}(1 \leqq i \leqq I_n, 1 \leqq j \leqq J_n)$ be independent random variables with zero means and finite variances

$$\operatorname{Var}(X_{i,j}^{(n)}) = a_i^{(n)}b_j^{(n)}$$

such that

$$\textstyle\sum_i a_i^{(n)} = 1 = \sum_j b_j^{(n)}.$$

Put

$$A_i^{(n)} = \textstyle\sum_{g\leqq i} a_g^{(n)} \qquad B_j^{(n)} = \sum_{h\leqq j} b_h^{(n)},$$

and define D_2-valued processes S_n by

$$S_n(t) = \textstyle\sum_{i\leqq A^{(n)}(t)} \sum_{j\leqq B^{(n)}(t)} X_{i,j}^{(n)},$$

where $A^{(n)}(t)$ (resp. $B^{(n)}(t)$) is the largest $A_i^{(n)}$ (resp. $B_j^{(n)}$) not exceeding t_1 (resp. t_2) $(t = (t_1, t_2))$.

THEOREM 5. *If the* $X_{i,j}^{(n)}$ *satisfy Lindeberg's condition, namely*

$$\lim_n \left[\sum_i \sum_j \int_{[|X_{i,j}^{(n)}| \geqq \varepsilon]} (X_{i,j}^{(n)})^2 \, dP\right] = 0 \qquad \text{for all } \varepsilon > 0,$$

and if

$$\max_i a_i^{(n)} \to 0, \max_j b_j^{(n)} \to 0,$$

then $S_n \to S$, *where* S *is a Gaussian process with zero means and covariances*

$$\operatorname{Cov}(S(t_1, u_1), S(t_2, u_2)) = \min(t_1, t_2) \min(u_1, u_2)$$

(*i.e.* S *is a Brownian motion process on* $[0, 1]^2$).

PROOF. For each n, put $T^n = \{A_i^{(n)};\ 0 \leqq i \leqq I_n\} \times \{B_j^{(n)}:\ 0 \leqq j \leqq J_n\}$. If B and C are a pair of neighboring blocks with corner points in T^n, then by independence one has

$$E[S_n^2(B) S_n^2(C)] = \operatorname{Var}[S_n^2(B)] \operatorname{Var}[S_n^2(C)] = \lambda(B)\lambda(C)$$

where λ denotes Lebesgue measure on $[0, 1]^2$. Consequently inequality (3) holds for S_n with $\gamma_1 = 2 = \gamma_2$ and $\beta_1 = 1 = \beta_2$ (so that $\gamma = 4$, $\beta = 2 > 1$), and the theorem follows from the remarks after Theorem 3 (which is not itself directly applicable). □

L. LeCam informed us that in an unpublished work carried out several years ago, he used the methods of LeCam (1958) to prove a theorem, involving partial sum processes, which is analogous to the normal convergence criteria for sums of u.a.n. variables. This of course includes Theorem 4.

(II) *Sampling from finite populations.* Let $p_{1,N}, \cdots, p_{N,N}$ be N given points in $T = [0, 1]^q$. Suppose that m points are drawn at random without replacement from this population. Distribution free tests in the q-variate two sample problem involve comparing the distribution of the drawn points to that of the remaining ones (see Bickel (1969)). This is conveniently done in terms of the following process. Let H_N be the (non-random) distribution function of the uniform probability over $p_{1,N}, \cdots, p_{N,N}$, and let F_m (resp. G_n) be the (random) distribution function of the uniform measure over the m drawn (resp. $n = N - m$ undrawn) points. Define a D_q-valued process $X_{m,n}$ by

$$X_{m,n} = (mn/N)^{\frac{1}{2}}(F_m - G_n) = (mN/n)^{\frac{1}{2}}(F_m - H_N).$$

The convergence of the $X_{m,n}$ was studied by a different method in Bickel (1969); in particular, it was shown that if $H_N \to H$ as $N \to \infty$, then $X_{m,n} \to X$ as m and n tend to ∞, where X is a Gaussian process with zero means and covariances $\operatorname{Cov}(H(t), H(u)) = H(\min(t, u)) - H(t)H(u)$ (the minimum being computed coordinatewise). Here we show how Theorem 3 may be applied to establish the tightness condition (assuming H is continuous).

For any two neighboring blocks B and C in T, one has

$$E(X_{m,n}(B))^2 (X_{m,n}(C))^2 = (N/mn)^2 E(N_B - mp_B)^2 (N_C - mp_C)^2,$$

where N_B, N_C, and $N_D = N - N_B - N_C$ have a multiple hypergeometric distribution:

$$P\{N_B = i, N_C = j, N_D = k\} = \binom{Np_B}{i}\binom{Np_C}{j}\binom{Np_D}{k} \Big/ \binom{N}{m}$$

$(i+j+k = m)$ with $p_B = H_N(B), p_C = H_N(C), p_D = 1-p_B-p_C$. By the extended version of Theorem 3, it suffices to show that for $N \geqq 4$

$$(12) \qquad E(N_B - mp_B)^2(N_C - mp_C)^2 \leqq 33(mn/N)^2 H_N(B)H_N(C).$$

For this note that given N_B, N_C is hypergeometric with parameters Nq_B (total population size), Np_C (sub-population size), and $m-N_B$ (sample size) $(q_B = 1-p_B)$. It follows that the left-hand side of (12) does not exceed

$$(p_C/q_B)^2 E(N_B - mp_B)^4 + (p_C p_D/(q_B{}^2(Nq_B-1)))[mnq_B{}^2 E(N_B-mp_B)^2 \\ + q_B(m-n)E(N_B-mp_B)^3].$$

From David, Kendall and Barton (1966) page 216, one finds

$$E(N_B - mp_B)^4 \\ = \frac{[Np_Bq_B(p_B{}^3+q_B{}^3)(N(N+1)-6mn)mn + 3N^2p_B{}^2q_B{}^2mn(m-1)(n-1)]}{N(N-1)(N-2)(N-3)}$$

$$E(N_B-mp_B)^3 = Np_Bq_B(p_B-q_B)mn(n-m)/[N(N-1)(N-2)]$$

$$E(N_B-mp_B)^2 = mnp_Bq_B/(N-1).$$

Simple manipulations now yield (12).

(III) *Empirical distribution functions.* We lead into the next application of Theorem 3 with a central limit theorem for D_q-valued processes. Let $Z, Z_1, Z_2, \cdots$ be independent identically distributed D_q-valued processes. Suppose that Z vanishes along the lower boundary of $T = [0, 1]^q$, that $EZ(t) = 0$ for all t in T, and that there exists a continuous finite measure μ on T such that

$$EZ^2(B) \leqq \mu(B)$$

$$EZ^2(B)Z^2(C) \leqq \mu(B)\mu(C)$$

for all pairs of neighboring blocks B and C in T.

Define D_{q+1}-valued processes $X_n (n \geqq 1)$ by

$$X_n(s,t) = (n^{-\frac{1}{2}}) \textstyle\sum_{j \leqq [ns]} Z_j(t)$$

$(s \in I = [0, 1],\ t \in T)$. Suppose that there exists a D_q-valued Gaussian process $X = (X(s,t))_{s\in I,\, t\in T}$ with zero means and covariances

$$\mathrm{Cov}\,(X(s_1,t_1), X(s_2,t_2)) = \min\,(s_1,s_2)\Gamma(t_1,t_2),$$

where $\Gamma(t_1, t_2) = \mathrm{Cov}\,(Z(t_1), Z(t_2))$. Assume that X is almost surely continuous along the upper boundary of $I \times T$. Such an X exists by Theorem 4 if, for example, Γ is continuous.

THEOREM 6. *In the present context,* $X_n \to X$.

PROOF. This follows easily from the remark following Theorem 3, inequality (3), and the inequalities below:

(i) $n^{-2}E[\sum_{i<\alpha\leqq j} Z_\alpha(B)]^2[\sum_{j<\alpha\leqq k} Z_\alpha(B)]^2 = [(j-i)/n]\mu(B)\cdot[(k-j)/n]\mu(B)$

(ii) $n^{-2}E[\sum_{i<\alpha\leq j} Z_\alpha(B)]^2[\sum_{i<\alpha\leq j} Z_\alpha(C)]^2$

$$\leq n^{-2}[(j-i)EZ^2(B)Z^2(C)+(j-i)(j-i-1)EZ^2(B)EZ^2(C)$$
$$+2(j-i)(j-i-1)(E(Z(B)Z(C)))^2]$$
$$\leq 3[(j-i)/n]\mu(B)\cdot[(j-i)/n]\mu(C),$$

holding for neighboring blocks B and C in T. □

In passing, we note that it is known that for function space valued processes, even so simple a central limit theorem as the assertion $X_n(1,\cdot) \to X(1,\cdot)$ is not valid without assumptions beyond those of independence and equi-distribution, zero means, and finite variances (see, e.g. Dudley and Strassen (1969)). Now let $(U_k)_{k>1}$ be a sequence of i.i.d. T-valued random variables having a continuous distribution, say Q. Define D_q-valued processes Z_k by

$$Z_k(t) = I_{C(t)}(U_k) - Q(C(t)),$$

where $C(t) = \prod_p [0, t_p](t = (t_1, \cdots, t_q))$, and define G_k by

$$G_k(t) = (1/k^{\frac{1}{2}})\sum_{1\leq j\leq k} Z_k(t).$$

G_k is of course nothing but the normalized empirical distribution function based on $U_1, \cdots, U_k$. Define a D_{q+1}-valued process X_n by

$$X_n(s,t) = ([ns]/n)^{\frac{1}{2}} G_{[ns]}(t) = (n^{-\frac{1}{2}})\sum_{j\leq[ns]} Z_j(t)$$

($s \in [0, 1]$, $t \in T$). Since

$$EZ^2(B) = \mathrm{Var}(Z(B)) = Q(B)(1-Q(B)) \leq Q(B)$$
$$E(Z(B)Z(C))^2 = Q^2(B^c)Q^2(C)Q(B) + Q^2(B)Q^2(C^c)P(C)$$
$$+Q^2(B)Q^2(C)(1-Q(B)-Q(C))$$
$$\leq 3Q(B)Q(C),$$

Theorem 5 implies that the X_n converge weakly (to a Gaussian process having continuous sample paths). In particular, the $G_n = X_n(1,\cdot)$ converge weakly. But of course much more than this is true. For example, using the methods of Billingsley (1968), Section 17, one can easily deduce (cf. Wichura (1968)) that G_{N_n} converges, to the limit of the G_n's, whenever $(N_n)_{n\geq 1}$ is a sequence of positive, integer-valued random variables such that, for some sequence of constants $c_n \to \infty$, N_n/c_n converges in probability to a positive random variable (see Fernandez (1970) for a different approach).

REFERENCES

BICKEL, P. J. (1969). A distribution free version of the Smirnov test in the p-variate case. *Ann. Math. Statist.* **40** 1–23.

BILLINGSLEY, P. (1968). Convergence of probability measures. Wiley, New York.

CHENTSOV, N. N. (1956). Weak convergence of stochastic processes whose trajectories have no discontinuities of the second kind and the "heuristic" approach to the Kolmogorov–Smirnov tests. *Theor. Probability Appl.* **1** 140–144.

DAVID, F. N., KENDALL, M. G., and BARTON, D. E. (1966). *Symmetric Function and Allied Tables.* Cambridge Univ. Press.

DUDLEY, R. (1966). Weak convergence of probabilities on nonseparable metric spaces and empirical measures on Euclidean spaces. *Illinois J. Math.* **10** 109–126.

DUDLEY, R. and STRASSEN, V. (1969). The central limit theorem and ε-entropy. *Lecture Notes in Math.* **89** Springer–Verlag, Berlin, 224–231.

FERNANDEZ, P. (1970). On the weak convergence of random sums of independent random elements. Ph.D. dissertation, Univ. of California, Berkeley.

KUELBS, J. (1968). The invariance principle for a lattice of random variables. *Ann. Math. Statist.* **39** 382–389.

LECAM, L. (1957). Convergence in distribution of stochastic processes. *Univ. California Publ. Statist.* **2** 207–236.

LECAM, L. (1958). Remarques sur les variables aléatoires dans les espaces vectoriels non separables. *Publ. Inst. Stat. Univ. Paris* **7** 39–53.

NEUHAUS, G. (1969). Zur theorie der Konvergenz stochasticher Prozessen mit mehrdimensionalem Zeitparameter. Ph.D. dissertation, Westphälischen Wilhelms-Universität zu Münster.

NEVEU, J. (1965). *Mathematical Foundations of the Calculus of Probability.* Holden–Day, San Francisco.

PROHOROV, YU. V. (1956). Convergence of random processes and limit theorems in probability theory. *Theor. Probability Appl.* **1** 157–214.

PYKE, R. (1965). Spacings. *J. Roy. Statist. Soc. Ser. B* **27** 395–449.

PYKE, R. (1968). The weak convergence of the empirical process with random sample size. *Proc. Cambridge Philos. Soc.* **64** 155–160.

ROSÉN, B. (1964). Limit theorems for sampling from a finite population. *Ark. Math.* **5** 383–424.

SKOROHOD, A. V. (1956). Limit theorems for stochastic processes. *Theor. Probability Appl.* **1** 261–290.

STRAF, M. L. (1969). A general Skorohod space and its applications to the weak convergence of stochastic processes with several parameters. Ph.D. dissertation, Univ. of Chicago.

STRAF, M. L. (1970). Weak convergence of stochastic processes with several parameters. *Proc. Sixth Berkley Symp. Math. Statist. Prob.* (to appear).

WICHURA, M. J. (1968). On the weak convergence of non-Borel probabilities on a metric space. Ph.D. dissertation, Columbia Univ.

WICHURA, M. J. (1969). Inequalities with applications to the weak convergence of random processes with multidimensional time parameters. *Ann. Math. Statist.* **40** 681–687.

The Annals of Statistics
1981, Vol. 9, No. 6, 1301-1309

MINIMAX ESTIMATION OF THE MEAN OF A NORMAL DISTRIBUTION WHEN THE PARAMETER SPACE IS RESTRICTED[1]

BY P. J. BICKEL

University of California, Berkeley

If X is a $N(\theta, 1)$ random variable, let $\rho(m)$ be the minimax risk for estimation with quadratic loss subject to $|\theta| \le m$. Then $\rho(m) = 1 - \pi^2/m^2 + o(m^{-2})$. We exhibit estimates which are asymptotically minimax to this order as well as approximations to the least favorable prior distributions. The approximate least favorable distributions (correct to order m^{-2}) have density $m^{-1} \cos^2 \left(\frac{\pi}{2m} s\right)$, $|s| \le m$ rather than the naively expected uniform density on $[-m, m]$. We also show how our results extend to estimation of a vector mean and give some explicit solutions.

1. Introduction If we want to estimate the completely unknown mean of a normal distribution with known variance using quadratic loss, then the sample mean is, of course, minimax. If, however, we have prior knowledge that the mean lies in a known interval, say $[-m, m]$, then the sample mean is inadmissible and it is well known that the minimax estimate is Bayes with respect to a least favorable prior distribution concentrating on a finite number of points. For m small (≤ 1.05) Casella and Strawderman (1980) show that this distribution concentrates on the end points. As m increases, the number of points increases, their location and the masses assigned to them vary in an as yet unknown fashion so that as $m \to \infty$, the prior distributions approximate Lebesgue measure (conditionally) and the minimax risk tends to the variance of the sample mean.

In Section 2, we ascertain a little more precisely what the behavior of the minimax risk is for large m. We do this by rescaling the least favorable prior distributions to the interval $[-1, 1]$ and finding the limit of these rescaled distributions as the solution of the variational problem of minimizing Fisher information among distributions concentrating on $[-1, 1]$. We show that this limit has density $\cos^2 (\pi/2)\, x$, $|x| \le 1$ and deduce that (for sample size 1, variance 1) the minimax risk is $1 - \pi^2/m^2 + o(m^{-2})$ as $m \to \infty$.

The key idea in obtaining this result is an identity relating Bayes risk with respect to any prior distribution to Fisher information. This identity is implicit in Brown (1971) and is related to an identity of Stein (Hudson, 1978). This relation is also used in Bickel (1980) and was independently discovered and used by Marazzi (1980) as well as Levit (1980).

In Section 3 we extend these results to estimation in p dimensions. We obtain the expected qualitative break in the shape of the limits of the rescaled prior for $p \ge 3$ and, parenthetically, can deduce the inadmissibility of the sample mean for $p \ge 3$.

2. The one dimensional case. Let X be a random $N(\theta, 1)$ variable. Let $\mathscr{D} = \{$all estimates of $\theta\}$ and for $\delta \in \mathscr{D}$, define

$$R(\theta, \delta) = E_\theta(\delta - \theta)^2.$$

Received July, 1980; revised January, 1981.

[1] Research supported by Office of Naval Research Grant Number N00014 80 C 0163 and the Adolph and Mary Sprague Miller Foundation.

AMS 1980 *subject classifications*: 62F10, 62C99.

Key words and phrases. Minimax, estimation, Fisher information, James-Stein estimate.

If G is a Bayes prior probability distribution on R let $\delta(\cdot, G)$ denote the Bayes estimate and

$$r(G) = \int R(\theta, \delta(\cdot, G))\, G(d\theta)$$

denote its Bayes risk. The minimax risk for estimating θ, given that $|\theta| \le m$, is defined by

$$\rho(m) = \min \max\{R(\theta, \delta): |\theta| \le m, \delta \in \mathscr{D}\}.$$

It is well known by convexity and analyticity considerations that there is a unique symmetric Bayes prior distribution G_m^0 concentrating on a finite number of points such that $\delta(\cdot, G_m^0)$ is unique minimax and G_m^0 is least favorable. That is,

$$R(\theta, \delta(\cdot, G_m^0)) = \max\{R(\theta, \delta(\cdot, G_m^0)): |\theta| \le m\} = r(G_m^0) = \rho(m)$$

with G_m^0 probability 1.

The structure of G_m^0 has been studied for small m by Casella and Strawderman (1980) who showed that for $|m| \le 1.05$, G_m^0 assigns mass ½ each to $\pm m$. We proceed with our study of G_m^0 for m large.

Let G_1 be the distribution on $[-1, 1]$ with density

$$g_1(s) = \cos^2\left(\frac{\pi}{2} s\right), \; |s| \le 1$$
$$= 0 \text{ otherwise,}$$

and let G_m be the corresponding distribution scaled up to $[-m, m]$ with density given by,

$$g_m(s) = m^{-1} g_1(sm^{-1}).$$

Then $\{G_m\}$ are approximately least favorable in the following sense.

THEOREM 2.1: *As* $m \to \infty$

$$\rho(m) = r(G_m) + o(m^{-2}) \tag{2.1}$$

$$r(G_m) = 1 - \frac{\pi^2}{m^2} + o(m^{-2}). \tag{2.2}$$

Moreover, let $G_1^{(m)}$ be the distribution obtained by scaling G_m^0 down to $[-1, 1]$, *i.e.,*

$$G_1^{(m)}(s) = G_m^0(ms)$$

then

$$G_1^{(m)} \to G_1 \tag{2.3}$$

in the sense of weak convergence.

It is *not* true that $\delta(\cdot, G_m)$ are asymptotically minimax. In fact, $\limsup_m R(m, \delta(\cdot, G_m)) > 1$. However, asymptotically minimax estimates can be constructed as follows. Let

$$\bar{\psi}(x) = -\frac{g_1'}{g_1}(x) = \pi \tan\left(\frac{\pi}{2} x\right), \; |x| < 1. \tag{2.4}$$

Suppose $\{\psi_m\}$ is a sequence of functions and that $\{a_m\}$, $\{b_m\}$, $\{c_m\}$ are sequences of positive numbers with the following properties:

(a) $1 > a_m \downarrow 0$, $ma_m \to \infty$

(b) $\sup\{|\psi_m(x) - \bar{\psi}(x)| : |x| \le 1 - a_m^2\} \to 0$

(c) $\sup\{|\psi_m'(x) - \bar{\psi}'(x)| : |x| \le 1 - a_m^2\} \to 0$

(d) for $|x| \ge 1 - a_m^2$, $2|\psi_m'(x)| + \psi_m^2(x) \le b_m + c_m x^2$

(e) $b_m\{1 - \Phi(ma_m)\} \to 0$

(f) $c_m\{1 - \Phi(ma_m)\} \to 0$,

where Φ is the standard normal c.d.f. Let

$$n = m(1 - a_m)^{-1},$$

and define

$$\delta_m(x) = x - n^{-1}\psi_m(xn^{-1}).$$

THEOREM 2.2. *If properties (a) – (f) hold, the estimates $\{\delta_m\}$ are asymptotically optimal and have asymptotically constant risk on $[-m, m]$ in the sense that*

$$\max\left\{\left|R(\theta, \delta_m) - 1 + \frac{\pi^2}{m^2}\right| : |\theta| \le m\right\} = o(m^{-2}). \tag{2.5}$$

Estimates δ_m can readily be constructed. For example, let

$$\begin{aligned} \psi_m(x) &= \bar{\psi}(x), \ |x| \le 1 - a_m^2 \\ &= [\bar{\psi}(1 - a_m^2) + \bar{\psi}'(1 - a_m^2)\{x - (1 - a_m^2)\}] \operatorname{sgn} x, \ |x| > 1 - a_m^2. \end{aligned} \tag{2.6}$$

It is easy to see that we can then take $b_m \sim a_m^{-4}$, $c_m \sim a_m^{-8}$ and conditions (e), (f) reduce to

$$a_m^{-8}\{1 - \Phi(ma_m)\} \to 0.$$

It is also possible to establish

COROLLARY 2.1. *The estimates $\delta(\cdot, G_n)$ are optimal for n as above if*

$$ma_m^6 \to \infty.$$

PRELIMINARIES. The key to these theorems are two identities. The first is a special case of (13.4) of Brown (1971). For any prior distribution G let f_G be the density of the marginal distribution of X, ϕ the standard normal density.

$$f_G(x) = \phi * G(x) = \int_{-\infty}^{\infty} \phi(x - \theta) G(d\theta).$$

(Here and in the sequel * denotes convolution.) Brown's identity for the Bayes risk of G is

$$r(G) = 1 - \int_{-\infty}^{\infty} \frac{\{f_G'(x)\}^2}{f_G(x)} dx. \tag{2.7}$$

The second identity is due to Stein, see Hudson (1978). Let δ be an estimate differentiable in x and such that

$$E_\theta|\delta'(X)| < \infty.$$

Let

$$\psi(x) = x - \delta(x).$$

Then Stein's identity is

$$R(\theta, \delta) = 1 - E_\theta\{2\psi'(X) - \psi^2(X)\}. \tag{2.8}$$

Stein's identity is obtained by an integration by parts while Brown's follows from Stein's by putting

$$\delta(x) = \delta(x, G) = x + \frac{f_G'(x)}{f_G(x)}$$

and integrating with respect to G. Brown's identity can be written in terms of the more familiar Fisher information defined (Huber, 1964) for any distribution F by

$$I(F) = \int_{-\infty}^{\infty} \frac{\{f'(x)\}^2}{f(x)}\,dx$$

if F has an absolutely continuous density f, being infinite otherwise. Evidently (2.7) is just

$$r(G) = 1 - I(\Phi * G).$$

We need four properties of $I(\cdot)$ which may be found in Port and Stone (1974).

(i) If $H(x) = F\left(\frac{x-\mu}{\sigma}\right)$, all x and $\sigma > 0$, then $I(H) = \sigma^{-2} I(F)$;

(ii) If $F_n \to F$ weakly, then $I(F) \le \liminf_n I(F_n)$;

(iii) If $H_n \to \delta_0$ (point mass at 0) weakly then $I(F * H_n) \to I(F)$;

(iv) $I(F_1 * F_2) \le \max\{I(F_1), I(F_2)\}$.

Finally, we require a special case of a theorem of Huber (1974).

LEMMA 2.1. *The distribution G_1 minimizes $I(F)$ uniquely among all F concentrating on* $[-1, 1]$. *Moreover,*

$$I(G_1) = \pi^2. \tag{2.9}$$

This follows (after an obvious typographical correction) from Huber's work since G_1 does concentrate on $[-1, 1]$ and is of the right form, i.e.,

$$\frac{(g_1^{1/2})''}{g_1^{1/2}} = \frac{1}{4}\left\{\frac{2g_1''}{g_1} - \left(\frac{g_1'}{g_1}\right)^2\right\} = \frac{-\pi^2}{4}. \tag{2.10}$$

We can now see where Theorem 2.1 comes from. By Brown's identity (2.7) we have

$$\begin{aligned}\rho(m) &= \sup\{r(G) : G \text{ concentrating on } [-m, m]\}\\ &= 1 - \inf\{I(\Phi * G) : G \text{ concentrating on } [-m, m]\}\end{aligned}$$

which by property (i) of I is then equal to

$$1 - m^{-2} \inf\{I(\Phi_{1/m} * G) : G \text{ concentrating on } [-1, 1]\}$$

where Φ_σ is the $N(0, \sigma^2)$ c.d.f. By Lemma 2.1, the coefficient of m^{-2} should be approximately $I(G_1)$ for m large. Here is a formal proof.

PROOF. Since

$$r(G_m) = 1 - I(\Phi * G_m) = 1 - m^{-2} I(\Phi_{1/m} * G_1)$$

(2.2) follows from property (iii) of I. Since G_m^0 is least favorable

$$\rho(m) = r(G_m^0) = 1 - I(\Phi * G_m^0) = 1 - m^{-2} I(\Phi_{1/m} * G_1^{(m)}). \tag{2.11}$$

Suppose (without loss of generality, since $[-1, 1]$ is compact) that $G_1^{(m)} \to G$ weakly. Then so does $\Phi_{1/m} * G_1^{(m)}$ and by property (ii) of I and (2.11) we must have

$$I(G) \le \liminf_m m^2\{1 - \rho(m)\}.$$

On the other hand, by property (iv) of I,

$$m^2\{1 - \rho(m)\} \le m^2\{1 - \rho(G_m)\} = I(\Phi_{1/m} * G_1) \le I(G_1).$$

and integrating with respect to G. Brown's identity can be written in terms of the more familiar Fisher information defined (Huber, 1964) for any distribution F by

$$I(F) = \int_{-\infty}^{\infty} \frac{\{f'(x)\}^2}{f(x)} dx$$

if F has an absolutely continuous density f, being infinite otherwise. Evidently (2.7) is just

$$r(G) = 1 - I(\Phi * G).$$

We need four properties of $I(\cdot)$ which may be found in Port and Stone (1974).

(i) If $H(x) = F\left(\frac{x-\mu}{\sigma}\right)$, all x and $\sigma > 0$, then $I(H) = \sigma^{-2} I(F)$;

(ii) If $F_n \to F$ weakly, then $I(F) \le \liminf_n I(F_n)$;

(iii) If $H_n \to \delta_0$ (point mass at 0) weakly then $I(F * H_n) \to I(F)$;

(iv) $I(F_1 * F_2) \le \max\{I(F_1), I(F_2)\}$.

Finally, we require a special case of a theorem of Huber (1974).

LEMMA 2.1. *The distribution G_1 minimizes $I(F)$ uniquely among all F concentrating on* $[-1, 1]$. *Moreover,*

$$I(G_1) = \pi^2. \tag{2.9}$$

This follows (after an obvious typographical correction) from Huber's work since G_1 does concentrate on $[-1, 1]$ and is of the right form, i.e.,

$$\frac{(g_1^{1/2})''}{g_1^{1/2}} = \frac{1}{4}\left\{\frac{2g_1''}{g_1} - \left(\frac{g_1'}{g_1}\right)^2\right\} = \frac{-\pi^2}{4}. \tag{2.10}$$

We can now see where Theorem 2.1 comes from. By Brown's identity (2.7) we have

$$\begin{aligned}\rho(m) &= \sup\{r(G) : G \text{ concentrating on } [-m, m]\}\\ &= 1 - \inf\{I(\Phi * G) : G \text{ concentrating on } [-m, m]\}\end{aligned}$$

which by property (i) of I is then equal to

$$1 - m^{-2} \inf\{I(\Phi_{1/m} * G) : G \text{ concentrating on } [-1, 1]\}$$

where Φ_σ is the $N(0, \sigma^2)$ c.d.f. By Lemma 2.1, the coefficient of m^{-2} should be approximately $I(G_1)$ for m large. Here is a formal proof.

PROOF. Since

$$r(G_m) = 1 - I(\Phi * G_m) = 1 - m^{-2} I(\Phi_{1/m} * G_1)$$

(2.2) follows from property (iii) of I. Since G_m^0 is least favorable

$$\rho(m) = r(G_m^0) = 1 - I(\Phi * G_m^0) = 1 - m^{-2} I(\Phi_{1/m} * G_1^{(m)}). \tag{2.11}$$

Suppose (without loss of generality, since $[-1, 1]$ is compact) that $G_1^{(m)} \to G$ weakly. Then so does $\Phi_{1/m} * G_1^{(m)}$ and by property (ii) of I and (2.11) we must have

$$I(G) \le \liminf_m m^2\{1 - \rho(m)\}.$$

On the other hand, by property (iv) of I,

$$m^2\{1 - \rho(m)\} \le m^2\{1 - \rho(G_m)\} = I(\Phi_{1/m} * G_1) \le I(G_1).$$

concentrating on $[-n, n]$ and $(\partial/\partial x)\,\delta(x, G_n)$ is its variance. Hence, by (2.14)

$$\left|\frac{h_n'(x)}{h_n(x)}\right| \le n^2(x+1) \qquad \left|\frac{h_n''(x)}{h_n(x)} - \left\{\frac{h_n'(x)}{h_n(x)}\right\}^2\right| \le n^2(n^2+1).$$

Therefore condition (d) holds, and from these estimates it is easy to see that conditions (e) and (f) are also satisfied if $ma_m^6 \to \infty$. □

3. The p variate case. Suppose X is $N_p(\theta, I)$ and that we want to estimate θ with quadratic loss. Then, the risk of an estimate δ is

$$R(\theta, \delta) = E_\theta \| \delta - \theta \|^2,$$

where $\|\ \|$ is the Euclidean distance. Conserving our previous notation, we consider the minimax risk of estimation given that $\|\theta\| \le m$, defined by $\rho_p(m) = \min \max\{R(\theta, \delta) : \|\theta\| \le m, \delta \in \mathscr{D}\}$.

By invariance the minimax estimate is Bayes with respect to a unique spherically symmetric least favorable prior distribution G_{mp}^0 concentrating (by analyticity considerations) on a finite number of spherical shells. We can again approximate G_{mp}^0 for large m.

Let J_t be the Bessel function of the first kind of order t, see Erdelyi et al. (1953), and let γ_t be its first positive zero. Let G_{1p} be the spherically symmetric distribution on the unit sphere $\{\theta : \|\theta\| \le 1\}$ with density given by

$$\begin{aligned} g_{1p}(\|x\|) &= C_p \|x\|^{-2t} J_t^2(\|x\|\gamma_t), && \|x\| \le 1, \\ &= 0, && \|x\| > 1, \end{aligned}$$

where

$$\begin{aligned} t &= \frac{p}{2} - 1 && \text{if } p \text{ is odd or divisible by 4} \\ &= -\left(\frac{p}{2} - 1\right) && \text{if } p \text{ is even and not divisible by 4} \end{aligned}$$

and c_p normalizes the density. It is well known that J_t has positive zeros (ibid, page 59) and by the standard representations (ibid, pages 2, 6) that $g_{1p}(0) > 0$. Moreover, $g_{1p}(r)$ is twice continuously differentiable on $[0, 1]$ and

(3.1) $$g_{1p}'(0) = g_{1p}'(1) = 0.$$

Let $$G_{mp}(s) = G_{1p}\left(\frac{s}{m}\right).$$

The generalization of Theorem 2.1 is then as follows.

THEOREM 3.1. *As $m \to \infty$,*

(3.2) $$\rho_p(m) = r(G_{mp}) + o(m^{-2})$$

(3.3) $$r(G_{mp}) = p - 4\gamma_t^2 m^{-2} + o(m^{-2});$$

and if $G_{1p}^{(m)}(s) = G_{mp}^{(0)}(ms)$ then

(3.4) $$G_{1p}^{(m)} \to G_{1p}$$

weakly as $m \to \infty$.

An analogue of Theorem 2.2 also holds. For simplicity we give the simplest example of asymptotically optimal estimates. Let

$$\bar{\psi}_p(r) = -\frac{g_{1p}'(r)}{g_{1p}(r)} = -\left\{2\gamma_t \frac{J_t'(\gamma_t r)}{J_t(\gamma_t r)} - \frac{(p-2)}{r}\right\}, \qquad r \ge 0.$$

$$\delta_m(x) = \left\{1 - n^{-1}\psi_{mp}\left(\frac{\| x \|}{n}\right) \| x \|^{-1}\right\}x,$$

where

$$\begin{aligned} \psi_{mp}(r) &= \bar{\psi}_p(r), & 0 \le r \le 1 - a_t^2 \\ &= \bar{\psi}_p(1 - a_t^2) + \bar{\psi}_p'(1 - a_t^2)\{r - (1 - a_t^2)\}, & r > 1 - a_t^2 \end{aligned}$$

and $n = m(1 + a_m)$. Then

$$\sup\left\{\left| R(\theta, \delta_m) - p + \frac{4\gamma_t^2}{m^2}\right| : \| \theta \| < m\right\} = o(m^{-2}) \tag{3.5}$$

if, for instance, $a_m \sim m^{-\epsilon}$, $0 < \epsilon < 1$.

NOTES:

1) As $m \to \infty$, $\forall x$, $\delta_{mp}(x) \to \left(1 - \frac{(p-2)}{\| x \|^2}\right)x$, Stein's (inadmissible) improvement for $p > 2$. Minimaxity of Stein's estimate follows.
2) These solutions have, for odd p, representations in terms of trigonometric and rational functions (Whittaker and Watson, 1927, page 364). In particular, for $p = 3$,

$$\begin{aligned} g_{13}(r) &= \frac{1}{2\pi}\frac{\sin^2(\pi r)}{r^2}, & 0 \le r \le 1 \\ &= 0 & \text{otherwise,} \end{aligned}$$

and correspondingly

$$\gamma_{1/2} = \pi$$

which can be contrasted with the value $\gamma_{-1/2} = \frac{1}{2}\pi$ for $p = 1$.

These results are based on the general forms of Brown's and Stein's identities, which we give in the following form. Let $I(F)$ be the Fisher information for the p-variate location problem as defined for instance in Port and Stone (1974). If F has a density f with continuous partial derivatives

$$I(F) = \int_{R^p} \left\{\Sigma_{j=1}^p \left(\frac{\partial f}{\partial x_j}\right)^2 (x)\right\} f^{-1}(x)\, dx.$$

Let

$$\delta(x) = x - \psi(x), \qquad \psi = (\psi_1, \cdots, \psi_p)$$

where $E_\theta \left| \frac{\partial}{\partial x_j} \psi_j(X) \right| < \infty$, $j = 1, \cdots, p$. Then

Brown's identity: For any prior distribution G,

$$r(G) = p - I(G * \Phi). \tag{3.6}$$

Stein's identity:

$$R(\theta, \delta) = p - E_\theta\left\{2\Sigma_{j=1}^p \frac{\partial}{\partial x_j}\psi_j(X) - \Sigma_{j=1}^p \psi_j^2(X)\right\}. \tag{3.7}$$

The generalization of Lemma 2.1 needed is

LEMMA 3.1. *The distribution G_{1p} uniquely minimizes $I(F)$ among all spherically symmetric F concentrating on the unit sphere and*

$$I(G_{1p}) = 4\gamma_t^2. \tag{3.8}$$

Moreover, $\sqrt{g_{1p}}$ *on* (0, 1) *satisfies the equation*

$$u''(r) + \frac{(p-1)}{r} u'(r) = -\gamma_t^2 u(r), \tag{3.9}$$

or equivalently

$$2\frac{g''_{1p}}{g_{1p}} - \left(\frac{g'_{1p}}{g_{1p}}\right)^2 + 2\frac{(p-1)}{r}\frac{g'_{1p}}{g_{1p}} = -4\gamma_t^2. \tag{3.10}$$

We can prove this lemma as in Huber (1974) (see also Huber, 1977) by considering the equivalent variational problem of minimizing $\int_0^\infty r^{p-1}\frac{\{f'(r)\}^2}{f(r)}\,dr$ subject to $\int_0^\infty r^{p-1}f(r)\,dr$ = constant. Equation (3.10) is equivalent to the Euler equation for the associated Lagrange problem of minimizing

$$\int_0^1 r^{p-1}\frac{\{f'(r)\}^2}{f(r)}\,dr - 4\gamma_t^2\int_0^1 r^{p-1}f(r)\,dr.$$

Convexity of the functional guarantees that a smooth solution of the Euler equation which satisfies the side conditions achieves the minimum. Unicity of a solution which is positive on (0, 1) is argued as in Huber (1974). Relation (3.8) follows by integrating (3.9) with respect to g and applying the identity

$$g''_{1p}(\|x\|) + \frac{(p-1)}{\|x\|}g'_{1p}(\|x\|) = \Sigma_{j=1}^p \frac{\partial^2}{\partial x_j^2} g_{1p}(\|x\|),$$

Gauss' theorem (e.g., Courant, 1937, page 401–402) and (3.1).

Claim (3.5) and similar results follow as in the one dimensional case when we note that if $\psi(x) = xw(\|x\|)/\|x\|$, where w is a smooth scalar function, then Brown's identity becomes

$$R(\theta,\delta) = p - E_\theta\left[2\left\{w'(\|x\|) + \frac{(p-1)}{\|x\|}w(\|x\|) - w^2(\|x\|)\right\}\right].$$

Generalizations in a variety of directions are possible. For example:
(1) to loss functions $l(\theta, \delta) = \Sigma_{j=1}^p \lambda_j(\delta_j - \theta_j)^2$ or equivalently to the case where X is $N(\theta, D)$ with D diagonal, known. Unfortunately Euler's equation now becomes a general elliptic partial differential equation.
(2) to study of the minimax risk over other sequences of growing regions. In general, this problem also seems very difficult. However, we note an interesting special case. The minimax risk over $\{\theta : \max_j|\theta_j| \le m\}$ is $p - p(\pi^2/m^2) + o(m^{-2})$ and is obtainable by using an asymptotically minimax estimate for each coordinate separately.

Acknowledgment. After submission of this paper I learned that B. Ya. Levit had just published and previously announced more extensive results of the same type (in Russian) in Levit (1980a, b). Further work is announced in Levit (1980c, d). The methods in this paper are somewhat different and perhaps simpler than Levit's.

REFERENCES

BICKEL, P. J. (1980). Minimax estimation of the mean of a normal distribution subject to doing well at a point. Unpublished technical report, Univ. of California, Berkeley.

BROWN, L. D. (1971). Admissible estimators, recurrent diffusions and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–903.

CASELLA, G. and STRAWDERMAN, W. (1981). Estimating a bounded normal mean. *Ann. Statist.* **9** 868–876.

COURANT, R. (1937). *Differential and Integral Calculus*, V. II. Interscience, New York.

ERDELYI, A. (1953). *Higher Transcendental Functions*, V. II. McGraw-Hill, New York.

GHOSH, M. N. (1964). Uniform approximation of minimax point estimates. *Ann. Math. Statist.* **35** 1031–1047.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

HUBER, P. J. (1974). Fisher information and spline interpolation. *Ann. Statist.* **2** 1029–1034.

HUBER, P. J. (1977). Robust covariances. *Statistical Decision Theory and Related Topics* II. Academic, New York.

HUDSON, H. M. (1978). A natural identity for exponential families with applications in multiparameter estimation. *Ann. Statist.* **6** 473–484.

LEVIT, B. YA. (1980a). On the second order asymptotically minimax estimates. *Theory Probab. Appl.* **25** 561–576.

LEVIT, B. YA. and BERHIN, P. E. (1980b). Second order asymptotically minimax estimates of the mean of a normal distribution. *Problems Inform. Transmission* **16** 60–79.

LEVIT, B. YA. (1980c). On the second order optimality in estimation theory. *Theory Probab. Appl.* **25** 653–654.

LEVIT, B. YA. (1980d). On some new results in the theory of second order optimality. *Theory Probab. Appl.* **25** 669–670.

MARAZZI, A. (1980). Robust Bayesian estimation for the linear model. Unpublished technical report, Fachgruppe fur Statistik, E. T. H. Zürich.

PORT, S. and STONE, C. (1974). Fisher information and the Pitman estimation of a location parameter. *Ann. Statist.* **2** 225–247.

WHITTAKER, E. T. and WATSON, G. N. (1927). *A Course of Modern Analysis.* Cambridge University Press, Cambridge.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720

The Annals of Probability
1983, Vol. 11, No. 1, 185–214

SUMS OF FUNCTIONS OF NEAREST NEIGHBOR DISTANCES, MOMENT BOUNDS, LIMIT THEOREMS AND A GOODNESS OF FIT TEST

BY PETER J. BICKEL[1] AND LEO BREIMAN[2]

University of California, Berkeley

We study the limiting behavior of sums of functions of nearest neighbor distances for an m dimensional sample. We establish a central limit theorem and moment bounds for such sums and an invariance principle for the empirical process of nearest neighbor distances. As a consequence we obtain the asymptotic behavior of a practicable goodness of fit test based on nearest neighbor distances.

1. Introduction and background. In many areas, there has been a long-standing need for a multidimensional goodness-of-fit test that is general, in the sense that the χ^2 and Kolmogorov-Smirnov test are general in one dimension, and also, is practical in a computational sense. Of course, χ^2 is still available in any number of dimensions, but its usefulness and practicality are virtually nil in high-dimensional spaces.

Take $X_1, \cdots, X_n$ to be n points in m-dimensional Euclidean space selected independently from a distribution with density $f(x)$. Define the nearest neighbor distance R_{jn} from X_j as

$$R_{jn} = \min_{1 \le i \ne j \le n} \|X_i - X_j\|.$$

In what follows we suppress the dependence of R_{jn} and related quantities on n unless confusion is likely.

The distance $d(x, y)$ between points does not have to be Euclidean. But we assume that it is generated by a norm $\|x\|$, i.e. $d(x, y) = \|x - y\|$.

This paper started with the attempt to derive the limiting distribution of a goodness of fit test for multidimensional densities based on the nearest neighbor distances. We established a form of the invariance principle. Our work had two main byproducts: a central limit theorem for sums of functions of nearest neighbor distances and 4th order moment bounds. These two pieces were then put together to get the invariance result.

The goodness of fit test. In looking for a practical goodness-of-fit test applicable to densities in an arbitrary number of dimensions, our starting point was the observation, essentially contained in the work by Loftsgaarden and Quesenberry (1965) that the variables

$$U_{jn} = \exp\left[-n \int_{\|x - X_j\| < R_j} f(x)\, d\mathbf{x}\right], \qquad j = 1, \cdots, n,$$

where $f(x)$ is the underlying density, $X_1, \cdots, X_n$, are n points sampled independently from $f(x)$ and R_j is the distance from X_j to its nearest neighbor, have a univariate distribution that, in any norm $\|\cdot\|$ distance

a; does not depend on $f(x)$

b; is approximately uniform.

The reasoning is simple: let $S(x, r)$ be the sphere with center at x and radius r. For any Borel set A, denote

Received November 1980; revised February 1982.

[1] This research was supported in part by contracts ONR N00014-75-C-0444 and N00014-80-C-0163.

[2] This research was supported in part by ONR Contract N00014-82-K-0054.

AMS 1980 *subject classifications.* 60F05, 62G10.

Key words and phrases. Nearest neighbor distances, goodness of fit, asymptotics.

$$F(A) = \int_A f(y)\,dy.$$

Assume X_1 is the first point selected, then the other $n - 1$. The set $\{R_1 \geq r_1\}$ is equal to the event that none of the $X_2, \cdots, X_n$ fall in the interior of the sphere of radius r_1 about X_1. Hence

$$P(R_1 \geq r_1 \mid X_1 = x_1) = [1 - F(S(x_1, r_1))]^{n-1}.$$

Since for fixed x, $F(S(x, r))$ is monotonically nondecreasing in r, write the above as

$$P[F(S(R_1, x_1)) \geq F(S(r_1, x_1)) \mid X_1 = x_1] = [1 - F(S(r_1, x_1))]^{n-1}.$$

Substituting $z = F(S(x_1, r_1))$ gives

$$P[F(S(x_1, R_1)) \geq z \mid X_1 = x_1] = (1 - z)^{n-1} \tag{1.1}$$

so that

$$P[F(S(X_1, R_1)) \geq z] = (1 - z)^{n-1}.$$

Since

$$U_1 = \exp[-nF(S(X_1, R_1))],$$

we have that for $\log x > -n$,

$$P(U_1 \leq x) = (1 + 1/n \log x)^{n-1} \sim x, \quad \text{for } x \text{ fixed}.$$

The above suggests that a possible approach to a goodness-of-fit test would be to take the density $g(x)$ to be tested, compute the statistics

$$\exp\left[-n \int_{S(X_j,R_j)} g(x)\,d\mathbf{x}\right]$$

and see whether, in some sense, the cumulative distribution function of these n variables is close to the uniform. While this is attractive theoretically, the computations involved in integrating anything but a very simple density over m-dimensional spheres are usually not feasible.

We reasoned that for n large, the nearest neighbor distances were small, on the average, and hence that we could use the approximation

$$\int_{S(X_j,R_j)} g(x)\,dx \sim g(X_j)V(R_j)$$

where

$$V(r) = K_m r^m$$

is the volume of an m-dimensional sphere of radius r. In this way we were led to testing based on the variables

$$W_j = \exp[-ng(X_j)V(R_j)], \qquad j = 1, \cdots, n.$$

An example of a measure of deviation of the W_j variables from the uniform is the statistic

$$S = \textstyle\sum_1^n (W_{(j)} - j/n)^2$$

where $W_{(j)}, j = 1, \cdots, n$, are the ordered W_j variables. Notice that

$$S = n \int_0^1 (\hat{H}(x) - x)^2\,d\hat{H}(x)$$

where $\hat{H}(x)$ is the sample d.f. of the W_j.

The invariance principle. This leads us more generally to studying the stochastic process $\hat{H}(y)$: $0 \leq y \leq 1$, and test statistics based on measures of the deviation of $\hat{H}$ from the uniform or, more appropriately, on the deviations of $\hat{H}$ from its expectation $E\hat{H}$. We had conjectured, based on some simulation studies, that statistics such as S were asymptotically distribution free under the null hypothesis. More generally, we had conjectured that the limiting distribution of $\sqrt{n}\,(\hat{H}(t) - t)$ was a Gaussian process with zero mean and a covariance not depending on $f(x)$. Our main result, as given in Section 5, is that this is almost true. What holds is that for the sequence of processes

$$Z_n(t) = \sqrt{n}(\hat{H}(t) - E\hat{H}(t)), \qquad Z_n \to_w Z$$

where $Z(t)$, $0 \leq t \leq 1$, is a zero mean Gaussian process whose covariance depends on the hypothesized density g and true density f, and indeed if $g = f$, then the covariance does not depend on f. The proof of this theorem and other results related to the goodness-of-fit test are given in Section 5.

Defining variables D_{jn} by

$$D_{jn} = n^{1/m} R_{jn},$$

then W_{jn} has the form

$$W_j = \phi(X_j, D_{jn})$$

and, denoting the indicator function by $I(\cdot)$,

$$\begin{aligned} Z_n(t) = \sqrt{n}(\hat{H}(t) - E\hat{H}(t)) &= \frac{1}{\sqrt{n}} \textstyle\sum_1^n [I(W_j \leq t) - EI(W_j \leq t)] \\ &= \frac{1}{\sqrt{n}} \textstyle\sum_1^n [h(X_j, D_j) - Eh(X_j, D_j)] \end{aligned}$$

for an appropriate h.

This identification suggests that the appropriate tools for the invariance principle are a central limit theorem and moment bounds and convergence theorems for sums of functions of nearest neighbor distances.

A central limit theorem. The central limit result established in Sections 3 and 4 is that for a function $h(x, d)$ on $E^{(m)} \times [0, \infty) \to E^{(1)}$ such that h is uniformly bounded and almost everywhere continuous with respect to Lebesgue measure,

$$\operatorname{Var}\left(\frac{1}{\sqrt{n}} \textstyle\sum_1^n h(X_j, D_j)\right) \to \sigma^2 < \infty$$

and

$$\frac{1}{\sqrt{n}} \textstyle\sum_1^n h^*(X_j, D_j) \to_{\mathscr{D}} N(0, \sigma^2)$$

where we make the convention here and through the rest of the paper that for any function $h(X_j, D_j)$

$$h^*(X_j, D_j) = h(X_j, D_j) - Eh(X_j, D_j).$$

This is generalized to a multidimensional central limit theorem, and used to give the result that

$$(Z_n(t_1), \cdots, Z_n(t_k)) \to_{\mathscr{D}} (Z(t_1), \cdots, Z(t_k)).$$

Our proof is long. We believe that this is due to the complexity of the problem. Nearest neighbor distances are not independent. But for large sample size the nearest neighbor distance to a point in one region of space is "almost" independent of the nearest neighbor distances in another region of space. The main idea for capitalizing on this large scale independence is to cut the space into a finite number of cells. For any point in a given cell, let its revised nearest neighbor distance be defined using only its neighbors in the *same* cell. The first step, then, is to show that asymptotically the revised nearest neighbor distances can be substituted for the original nearest neighbor distances. Now, given the number of points in each cell, the set of interpoint distances within the Jth cell is independent of those within any other cell. Therefore, given the total cell populations, any sum of functions of the revised nearest neighbor distances is a sum of independent components, with each such component being the sum of the functions of the nearest neighbor distances within a particular cell.

However, the multinomial fluctuation of the cell population is *not* asymptotically negligible. Thus, the limiting distribution breaks into a sum of two parts, one being the nearly normal sum of the independent cell components given the expected value of the cell populations. The other is an asymptotically normal contribution due to the fluctuations of the cell populations from their expected values. The limiting form of the variance reflects the nature of the problem. It has one term that would be the variance if all nearest neighbor distances were assumed independent. Then there are a number of other, more complex, terms arising from the local dependence.

A moment bound. Both the central limit theorem and the tightness argument required for the invariance proof rely on moment bounds. Again, there is some difficulty in untangling the dependence between nearest neighbor distances and proving bounds of the type required.

For example, we show in Section 2 that for any measurable function h on $E^{(m)} \times [0, \infty) \to E^{(1)}$ with

$$\|h\| = \sup|h(x, d)| < \infty$$

there is a constant $M < \infty$ depending only, in a specified and useful way, on h and the dimension m such that

$$E(\sum h^*(X_j, D_j))^4 \le Mn^2.$$

Both the central limit theorem and the moment inequalities (which improve results in Rogers, 1977) should prove generally useful in methods employing nearest neighbor distances.

The plan of the presentation is

Section 2: moment bounds.
Section 3: 2nd moment convergence.
Section 4: central limit theorem.
Section 5: invariance and the goodness-of-fit test.
Appendix: technical results on nearest neighbor distances.

Section 2 on moment bounds is long and somewhat complex. But the results are needed in the later proofs. The main results of statistical interest are in Sections 4 and 5.

Assumptions on the densities. Our general assumptions on the density $f(x)$ are that it be uniformly bounded and continuous on its support. These requirements can probably be weakened, but the price may not be worth the extra generality. The following conditions are listed to make the requirements formal.

CONDITION A. We can choose a version of f such that
(i) $\{f > 0\}$ is open
(ii) f is continuous on $\{f > 0\}$
(iii) f is uniformly bounded.

Corresponding to A we have:

CONDITION B. The given function g is nonnegative and
(i) $\{g > 0\} \supset \{f > 0\}$
(ii) g is continuous on $\{f > 0\}$.

Clearly essentially all situations of interest are covered by A and B.

2. Some useful moment inequalities. The central result of this section is the 4th order moment bound (2.2) which is used to prove tightness via Corollary 2.5. We believe it will prove generally useful in the study of procedures based on nearest neighbors. Its formulation and spirit owe much to the excellent thesis of W. R. Rogers (1977). Our method of proof is, however, different from his and suited to the rather delicate estimates we must make.

The proof of the central limit theorem requires only the use of the 2nd order moment bounds given in Lemma 2.11 and its Corollary 2.15. The proofs of 2.11 and 2.15 are given early in this section and the reader interested only in the central limit problem may wish to skip the rest of the section.

The following notation is used:

P is the probability measure making $X_1, \cdots, X_n$ i.i.d. with common density f.
E without subscript is expectation under P.
R_i is the nearest neighbor distance to X_i.
J_i is the index of the nearest neighbor point to X_i.
$D_i = n^{1/m}R_i$
$I(A)$ is the indicator of an event.
$F(A) = \int_A f(y)\,dy$
$S(x, r) = \{y: \|y - x\| \le r\}$
$S_i = S(X_i, R_i)$

For h a measurable function on $E^{(m)} \times [0, \infty) \to E^{(1)}$, denote

$$\|h\| = \sup_{x,d} |h(x, d)|, \qquad h_i = h(X_i, D_i), \qquad h_i^* = h_i - Eh_i.$$

Throughout this section M, with or without a subscript, denotes a *finite* generic constant depending only on the dimension m.

THEOREM 2.1. *If* $\|h\| < \infty$, *then*

$$E(\textstyle\sum_i h_i^*)^4 \le Mn^2 \|h\|^2[E^2 |h_1| + n^4E^2 |h_1| F^2(S_1) + n^{-1} \|h\|^2]. \tag{2.2}$$

Before giving the proof of the theorem we give two corollaries.

COROLLARY 2.3. *Suppose u and w are bounded functions and*

$$h(x, d) = u(x)w(x, d).$$

Then there is a constant $C < \infty$ depending on $\|u\|$, $\|w\|$, m such that

$$E(\textstyle\sum_{i=1}^n h_i^*)^4 \le C(n^2E^2 |u(X_1)| + n). \tag{2.4}$$

PROOF. The corollary follows from

$$E |h_1| \le \|w\| E |u(X_1)|$$

$$E |h_1| F^2(S_1) \le \|w\| E\{E |u(X_1)| E(F^2(S_1) | X_1)\} = \|w\| E |u(X_1)| \frac{1}{n(n+1)}$$

where the last equality follows from (1.1).

COROLLARY 2.5. *If*

$$h(x, d) = I(a \le g(x)\, d^m \le b)$$

then

$$E\left(\textstyle\sum_i h_i^*\right)^4 \le M\{n^2(G_n(b) - G_n(a))^2 + n\} \tag{2.6}$$

where $G_n(y)$, $y \ge 0$, *is the distribution function defined by*

$$G_n(y) = \left(1 - \exp\left(-\frac{n}{2}\right)\right)^{-1} \int f(x)\left(1 - \exp\left[-\frac{n}{2} F(S(x, (y/ng(x))^{1/m}))\right]\right) dx.$$

PROOF. Let

$$\alpha(x) = F\left(S\left(x, \left(\frac{a}{ng(x)}\right)^{1/m}\right)\right)$$

$$\beta(x) = F\left(S\left(x, \left(\frac{b}{ng(x)}\right)^{1/m}\right)\right).$$

Then, for $j \ge 0$, defining $p_\alpha = F(S(x, \alpha))$, $p_\beta = F(S(x, \beta))$,

$$E(|h_1|\, F^j(S_1)\,|X_1 = x) = E[F^j(S(x, R_1)I(p_\alpha \le F(S(x, R_1)) \le p_\beta))\,|\,X_1 = x].$$

$$= \int_{p_\alpha}^{p_\beta} u^j(n-1)(1-u)^{n-2}\, du \le Mn^{-j} \int_{np_\alpha}^{np_\beta} w^j\left(1 - \frac{w}{n}\right)^{n-2} dw$$

or

$$E(|h_1|\, F^j(S_1)\,|X_1 = x) \le M_j n^{-j}\left(\exp\left(-\frac{np_\alpha}{2}\right) - \exp\left(-\frac{np_\beta}{2}\right)\right). \tag{2.7}$$

If we now apply Theorem 2.1 and use (2.7) for $j = 0, 1$ the lemma follows.

The proof of Theorem 2.1 proceeds by a construction similar to one used by Rogers and by a series of lemmas.

We assume that we are given a measurable set $S \subset R^m$, $F(S) < 1$, and a set of $r < n$ points, $\mathbf{x} = (x_1, \cdots, x_r)$, where the x_i are fixed points in S. Let $Q_r(\cdot \mid S, \mathbf{x})$ be the probability measure on $(R^m)^n$ such that $X_1, \cdots, X_{n-r}$ are independent identically distributed with their common distribution being the conditional distribution $F(\cdot \mid S^c)$ and $X_{n-r+1} = x_i$, $i = 1, \cdots, r$. We write $F(\cdot \mid S^c)$ as F_S. Its density is, of course,

$$\begin{aligned} f_S(x) &= f(x)/F(S^c), \qquad && x \in S^c \\ &= 0 && \text{otherwise.} \end{aligned}$$

We typically write Q_r for $Q_r(\cdot \mid S, \mathbf{x})$, and E_{Q_r} to denote the expectation under Q_r.

On a common probability space take $X_1, \cdots, X_n$ i.i.d. F and $Y_1, \cdots, Y_n$ i.i.d. $F(\cdot \mid S^c)$ and independent of the X_i and define,

$$\begin{aligned} \tilde{X}_i &= X_i \quad \text{if} \quad i = 1, \cdots, n-r \quad \text{and} \quad X_i \in S^c \\ &= Y_i \quad \text{if} \quad i = 1, \cdots, n-r \quad \text{and} \quad X_i \in S = x_{i-n+r} \quad \text{if} \quad i = n-r+1, \cdots, n. \end{aligned}$$

Clearly $\tilde{X}_1, \cdots, \tilde{X}_n$ have joint distribution Q_r. Let $\tilde{R}_i$ be the nearest neighbor distance of $\tilde{X}_i$ in the set $\tilde{X}_1, \cdots, \tilde{X}_n$ and $\tilde{D}_i$, $\tilde{J}_i$, $\tilde{S}_i$ be defined similarly.

LEMMA 2.8. *For* $n \ge r$, *there is a constant* M_0 *such that*

$$|E_{Q_r}h(X_1, D_1) - Eh(X_1, D_1)| \le \|h\|\, M_0\left(\frac{r}{n} + F(S)\right).$$

PROOF. For $r \geq n/2$, the bound holds trivially. For $n/2 > r$,

$$\begin{aligned} | E_{Q_r} h(X_1, D_1) - Eh(X_1, D_1)| &= (n-r)^{-1} \left| \textstyle\sum_{i=1}^{n-r} [Eh(X_i, D_i) - Eh(\tilde{X}_i, \tilde{D}_i)]\right| \\ &\leq (n-r)^{-1} E \textstyle\sum_{i=1}^{n-r} | h(X_i, D_i) - h(\tilde{X}_i, \tilde{D}_i)| \\ &\leq (n-r)^{-1} \| h \| E \textstyle\sum_{i=1}^{n-r} \{I(X_i \neq \tilde{X}_i) + I(X_i = \tilde{X}_i, R_i \neq \tilde{R}_i)\}. \end{aligned} \tag{2.9}$$

Let

$$N = \textstyle\sum_{i=1}^{n-r} I(X_i \neq \tilde{X}_i),$$

the number of "changed" points among the first $n - r$. Note that $EN = (n - r)F(S)$. Now

$$I(R_i \neq \tilde{R}_i, X_i = \tilde{X}_i) \leq \textstyle\sum_{j,k} I(J_i \neq j, \tilde{J}_i = k, X_j \neq \tilde{X}_j \quad \text{or} \quad X_k \neq \tilde{X}_k)$$

and hence

$$\begin{aligned} \textstyle\sum_i I(R_i \neq \tilde{R}_i, X_i = \tilde{X}_i) &\leq \textstyle\sum_j I(X_j \neq \tilde{X}_j) \textstyle\sum_i I(J_i = j) + \textstyle\sum_k I(X_k \neq \tilde{X}_k) \textstyle\sum_k I(J_i = k) \\ &\leq 2\alpha(m)(N + r) \end{aligned} \tag{2.10}$$

by Corollary S1 of the appendix.

From (2.9)–(2.10) and the boundedness of h,

$$\begin{aligned} | E_{Q_r} h_1 - Eh_1 | &\leq \| h \| \left\{ (1 + 2\alpha(m))F(S) + 2\alpha(m)\left(\frac{r}{n-r}\right)\right\} \\ &\leq \| h \| 2(1 + 2\alpha(m))\left(F(S) + \frac{r}{n}\right) \end{aligned}$$

and the lemma is proved.

LEMMA 2.11. *For $\|g\|$, $\|h\|$, $< \infty$, denote $h_1 = h(X_1, D_1)$, $g_2 = g(X_2, D_2)$. Then for $n \geq 4$,*

$$| \operatorname{cov}(h_1, g_2)| \leq M_1 \| g \|(n^{-1}E | h_1 | + E | h_1 F(S_1)|).$$

PROOF. Write

$$| \operatorname{cov}(h_1, g_2)| \leq \int_{[J_1=2]} | h_1^* g_2^* | \, dP + \left| \int_{[J_1 \neq 2]} h_1^* g_2^* \, dP \right|.$$

But

$$\int_{[J_1=2]} | h_i^* g_2^* \, dP | \leq \frac{2 \| g \|}{n-1} \textstyle\sum_{k=2}^{n} \int_{[J_1=k]} | h_1^* | \leq \frac{4 \| g \|}{n-1} E | h_1 |. \tag{2.12}$$

Moreover,

$$\int_{[J_1 \neq 2]} h_1^* g_2^* \, dP = \int_{[J_1 \neq 2]} h_1^* \{E(g_2 \mid X_1, X_{J_1}, J_1) - Eg_2\} \, dP. \tag{2.13}$$

On the set $J_1 \neq 2$, given $X_1 = x_1$, $X_{J_1} = x_2$, the $\{X_j, 2 \leq j \leq n, j \neq J_1; X_1, X_{J_1}\}$ are distributed according to $Q_2(\cdot \mid S(x_1, | x_2 - x_1|), (x_1, x_2))$. By Lemma 2.8

$$\begin{aligned} \left| \int_{[J_1 \neq 2]} h_1^* g_2^* \, dP \right| &\leq \int_{[J_1 \neq 2]} | h_1^* | M_0 \| g \|(2n^{-1} + F(S_1)) \, dP \\ &\leq 4M_0 \| g \| [n^{-1}E | h_1 + E | h_1 F(S_1)|] \end{aligned} \tag{2.14}$$

and the lemma follows from (2.12)–(2.14).

COROLLARY 2.15. *For $\|h\|, \|g\| < \infty$, and for $n \geq 4$,*

$$|\operatorname{cov}(h_1, g_2)| \leq M_2 \|g\| (Eh_1^2)^{1/2}/n.$$

PROOF. From (1.1) it follows that $EF^2(S_1) = 2/n(n+1)$. Now apply the Schwartz inequality.

The bounds in Lemma 2.11 and Corollary 2.15 can clearly be made symmetric in h_1 and g_2. We use them primarily for

LEMMA 2.16.

$$|\operatorname{cov}_{Q_r}(h_1, h_2) - \operatorname{cov}(h_1, h_2)| \leq \|h\|^2 M_3 \left(\frac{r^2}{n^2} + F^2(S)\right).$$

PROOF. Let $(X_1', \tilde{X}_1'), \cdots, (X_n', \tilde{X}_n')$ have the same joint distribution as the vector $\{(X_1, \tilde{X}_1), \cdots, (X_n, \tilde{X}_n)\}$ and be independent of that vector. Let primes on D_i, $\tilde{D}_i$, J_i, etc. as usual denote calculations based on the appropriate sample. Then

$$\operatorname{cov}(h_1, h_2) - \operatorname{cov}_{Q_r}(h_1, h_2) = \tfrac{1}{2} E\Delta \tag{2.17}$$
$$\Delta = (h_1 - h_1')(h_2 - h_2') - (\tilde{h}_1 - \tilde{h}_1')(\tilde{h}_2' - \tilde{h}_2')$$

where

$$h_i' = h(X_i', D_i'), \qquad \tilde{h}_i = h(\tilde{X}_i, \tilde{D}_i), \qquad \tilde{h}_i' = h(\tilde{X}_i', \tilde{D}_i').$$

The proof proceeds by a series of steps.

Let

$$E_i = \{h_i \neq \tilde{h}_i\}, \qquad E_i' = \{h_i' \neq \tilde{h}_i'\}.$$

Since

$$I(E_i) \leq I(X_i \neq \tilde{X}_i) + I(X_i = \tilde{X}_i, R_i \neq \tilde{R}_i),$$

Lemma A.1 and elementary arguments yield that

$$\max\{P(E_i \cap E_j), P(E_i \cap E_k') : \text{all } i, j, k, i \neq j\} \leq M\left(\frac{r^2}{n^2} + F^2(S)\right). \tag{2.18}$$

Since $\Delta = 0$ on $[U_{i=1}^2\{E_i U E_i'\}]^c$, (2.18) and symmetry arguments imply that

$$|E\Delta| \leq 4|E(h_1 - \tilde{h}_1)(h_2 - h_2')I(E_1 E_2^c [E_1']^c [E_2']^c)| + M\|h\|^2 \left(\frac{r^2}{n^2} + F^2(S)\right). \tag{2.19}$$

Using Lemma A.1 again we bound the first term on the right hand side of (2.19) by,

$$4|E\{(h_1 - \tilde{h}_1)(h_2 - h_2')(I(J_1 \neq 2, \tilde{J}_1 \neq 2, X_2 = \tilde{X}_2)[I(X_1 \neq \tilde{X}_1)$$
$$+ I(X_1 = \tilde{X}_1, R_1 \neq \tilde{R}_1)])\}| + M\|h\|^2\left(\frac{r^2}{n^2} + F^2(S)\right).$$

Let $\Xi = \{i : X_i \neq \tilde{X}_i\}$. Given Ξ, X_i, $i \in \Xi$, $X_1, X_{J_1}, \tilde{X}_1, X_{\tilde{J}_1}, \tilde{X}_{\tilde{J}_1}$ and $X_2 = \tilde{X}_2$ the variables $X_1, \cdots, X_n$ can be permuted to have a

$$Q_r(\cdot \mid S(X_1, R_1) U S(\tilde{X}_1, \tilde{R}_1), \{X_i, i \in \Xi, X_1, X_{J_1}, X_{\tilde{J}_1}\})$$

distribution with X_2 in the lead and $r = N + I(X_1 = \tilde{X}_1) + I(X_{J_1} = \tilde{X}_{J_1}) + I(X_{\tilde{J}_1} = \tilde{X}_{\tilde{J}_1})$. Conditioning on this information within the expectation in (2.20) and using the independence of h_2', we can apply Lemma 2.8 to the difference between the conditional expectation of h_2 and Eh_2' and bound the first term in (2.20) by

$$4\|h\|^2 M_0(m) E\left\{(I(X_1 \neq \tilde{X}_1) + I(X_1 = \tilde{X}_1, R_1 \neq \tilde{R}_1))\left(\frac{N+3}{n} + F(S_1) + F(\tilde{S}_1)\right)\right\}. \tag{2.21}$$

Estimates of the order $r^2/n^2 + F^2(S)$ for all the terms in (2.21) are given in Lemma A.2. Combining (2.19)–(2.21), the lemma follows.

LEMMA 2.22.

$$|Eh_1^*h_2^*h_3^*h_4^*| \le M_4\|h\|^2\left(\frac{E^2|h_1|}{n^2} + n^2E^2|h_1|F^2(S_1) + \|h\|^2n^{-3}\right). \tag{2.23}$$

PROOF. Let $E_{12} = [J_1, J_2 \not\in \{3, 4\}]$, $\pi = h_1^*h_2^*h_3^*h_4^*$.
Then,

$$\int_{E_{12}} \pi\, dP = \int_{E_{12}} h_1^*h_2^*\{\operatorname{cov}_{Q_r}(h_1, h_2) + (E_{Q_r}h_1 - Eh_1)^2\}\, dP \tag{2.24}$$

where

$$Q_r = Q_r(\cdot\,|\,S(X_1, R_1)\,US(X_2, R_2), \{X_1, X_2, X_{J_1}, X_{J_2}\}) \quad\text{and}\quad r \le 4.$$

Apply Lemmas 2.8, 2.11 and 2.16 to get,

$$\begin{aligned}\left|\int_{E_{12}} \pi\, dP\right| &\le (M_1\|h\|(n^{-1}E|h_1| + E|h_1F(S_1)|)) \times \left|\int_{E_{12}} h_1^*h_2^*\, dP\right| \\ &\quad + M_2\|h\|^2\int_{E_{12}} |h_1^*h_2^*|(n^{-2} + F^2(S_1) + F^2(S_2))\, dP.\end{aligned} \tag{2.25}$$

Next

$$\int_{E_{12}'} h_1^*h_2^* = 2\int_{[J_1=3]} h_1^*h_2^* + 2\int_{[J_2=3, J_1\not\in\{3,4\}]} h_1^*h_2^*. \tag{2.26}$$

Condition in the first integral on the right in (2.26) by X_1, X_{J_1}, J_1 and apply Lemma 2.8 to get the bound

$$2M_0\|h\|\int_{[J_1=3]} |h_1^*|(n^{-1} + F(S_1))\, dP \le 4M_0\frac{\|h\|}{n-1}(n^{-1}E|h_1| + E|h_1F(S_1)|) \tag{2.27}$$

by the usual symmetry argument. Condition in the second integral by X_2, X_{J_2}, J_2 and obtain a bound as in (2.27). Conclude that

$$\left|\int_{E_{12}} h_1^*h_2^*\right| \le |\operatorname{cov}(h_1, h_2)| + M\frac{\|h\|}{n}(E|h_1|n^{-1} + E|h_1F(S_1)|)$$

and hence that the first term in (2.25) is bounded by

$$M\|h\|^2\left(\frac{E^2|h_1|}{n^2} + E^2|h_1F(S_1)|\right). \tag{2.28}$$

On the other hand, applying Lemma 2.8 again

$$\begin{aligned}\int |h_1^*h_2^*|(n^{-2} + F^2(S_1)) &\le \|h\|\int_{[J_1=2]} |h_1^*|(n^{-2} + F^2(S_1)) \\ &\quad + \int_{[J_1\ne 2]} |h_1^*|(n^{-2} + F^2(S_1))\{E|h_2^*| + M_0\|h\|(n^{-1} + F(S_1))\}.\end{aligned} \tag{2.29}$$

The first term in (2.29) is $\le M\|h\|^2n^{-3}$ by the usual symmetry argument. The second is

$$\begin{aligned}&\le M((E^2|h_1|n^{-2} + E|h_1|E|h_1|F^2(S_1)) + \|h\|^2n^{-3}) \\ &\le M(2(E^2|h_1|n^{-2} + n^2E^2|h_1|F^2(S_1)) + \|h\|^2n^{-3})\end{aligned} \tag{2.30}$$

and hence combining (2.28) and (2.30) we get

$$\left| \int_{E_{12}} \pi \, dP \right| \le M \|h\|^2 \left(\frac{E^2|h_1|}{n^2} + E^2|h_1| F(S_1) + n^2 E^2 |h_1| F^2(S_1) + \|h\|^2 n^{-3} \right). \tag{2.31}$$

Now consider

$$\int_{[J_1=3]} \pi \, dP = \int_{[J_1=3, J_3 \notin \{2,4\}]} \pi \, dP + 2 \int_{[J_1=3, J_3=2]} \pi \, dP. \tag{2.32}$$

By conditioning on X_1, X_3, J_1, J_3, X_{J_1}, X_{J_3} we can bound the first integral on the right in (2.32) in exactly the same way as $\int_{E_{12}} \pi \, dP$ by

$$\begin{aligned} M \Bigg\{ & \|h\| \left(\frac{E|h_1|}{n} + E|h_1 F(S_1)| \right) \left| \int_{[J_1=3, J_3 \notin \{2,4\}]} h_1^* h_3^* \, dP \right| \\ & + \|h\|^2 \int_{[J_1=3]} |h_1^* h_3^*| (n^{-2} + F^2(S_1) + F^2(S_3)) \, dP \Bigg\}. \end{aligned} \tag{2.33}$$

Now use symmetry to bound

$$\left| \int_{[J_1=3, J_3 \notin \{2,4\}]} h_1^* h_3^* \right|$$

by

$$\frac{2\|h\|}{n-1} E|h_1|$$

and the second term in (2.33) by,

$$\frac{M\|h\|^4}{n^3}.$$

Hence,

$$\left| \int_{[J_1=3, J_3 \notin \{2,4\}]} \pi \, dP \right| \le M \|h\|^2 \left(\frac{E^2|h_1|}{n^2} + E^2 |h_1| F(S_1) + \|h\|^2 n^{-3} \right). \tag{2.34}$$

Next write,

$$\int_{[J_1=3, J_3=2]} \pi \, dP = \int_{[J_1=3, J_3=2, J_2 \ne 4]} \pi \, dP + \int_{[J_1=3, J_3=2, J_2=4]} \pi \, dP. \tag{2.35}$$

Now

$$\begin{aligned} P[J_1 = 3, J_3 = 2, J_2 = 4] &= \frac{1}{n-3} \textstyle\sum_{i=4}^{n} P[J_1 = 3, J_3 = 2, J_2 = i] \\ &\le (n-3)^{-1} P[J_1 = 3, J_3 = 2] \le (n-3)^{-1}(n-2)^{-1} P[J_1 = 3] \\ &\le M n^{-3}. \end{aligned} \tag{2.36}$$

Hence,

$$\left| \int_{[J_1=3, J_3=2, J_2=4]} \pi \, dP \right| \le M \|h\|^4 n^{-3}. \tag{2.37}$$

Next condition on X_1, X_2, X_3, J_1, J_2, J_3, R_1, R_2, R_3 in the first term of (2.35) and apply Lemma 2.8 to get

$$\left| \int_{[J_1=3, J_3=2, J_2 \ne 4]} \pi \, dP \right| \le M_0 \|h\|^4 \int_{[J_1=3, J_3=2]} (n^{-1} + \textstyle\sum_{i=1}^{3} F(S_i)) \, dP. \tag{2.38}$$

Now,

$$P[J_1 = 3, J_3 = 2] \leq Mn^{-2}$$

as in (2.36) and similarly,

$$\int_{[J_1=3, J_3=2]} F(S_1)\, dP \leq (n-2)^{-1} \int_{[J_1=3]} F(S_1)\, dP \tag{2.39}$$

$$= [(n-2)(n-1)]^{-1} EF(S_1) \leq Mn^{-3},$$

$$\int_{[J_1=3, J_3=2]} F(S_2)\, dP = (n-2)^{-1} \int_{[J_3=2]} F(S_2) \textstyle\sum_{i \neq 2,3} I(J_i = 3)\, dP \tag{2.40}$$

$$\leq (n-2)^{-1} \alpha(m) \int_{[J_3=2]} F(S_2)\, dP$$

by Corollary S1,

$$\leq [(n-2)(n-1)]^{-1} \alpha^2(m) \int F(S_2)\, dP \leq Mn^{-3}$$

$$\int_{[J_1=3, J_3=2]} F(S_3)\, dP \leq [(n-2)(n-1)]^{-1} \alpha(m) EF(S_3) \leq Mn^{-3}. \tag{2.41}$$

Combining these estimates with (2.38), (2.37) and (2.35) we get,

$$\left| \int_{[J_1=3, J_3=2]} \pi\, dP \right| \leq M \| h \|^4 n^{-3} \tag{2.42}$$

and hence from (2.32), (2.34) and (2.42),

$$\left| \int_{[J_1=3]} \pi\, dP \right| \leq M \| h \|^2 \left(\frac{E^2 | h_1 |}{n^2} + E^2 | h_1 | F(S_1) + \| h \|^2 n^{-3} \right). \tag{2.43}$$

Next consider,

$$\int_{[J_2=3, J_1 \notin \{3,4\}]} \pi\, dP = \int_{[J_2=3]} \pi\, dP - \int_{[J_1=J_2=3]} \pi\, dP - \int_{[J_2=3, J_1=4]} \pi\, dP. \tag{2.44}$$

Of these terms the first is bounded in (2.43). The next is written,

$$\int_{[J_1=J_2=3, J_3 \neq 4]} \pi\, dP + \int_{[J_1=J_2=3, J_3=4]} \pi\, dP. \tag{2.45}$$

The second term in (2.45) is bounded by $M \| h \|^4 n^{-3}$ as in (2.40). The first (conditioning on X_1, X_2, X_3, etc.) is bounded by

$$M \| h \|^4 \int_{[J_1=J_2=3]} (n^{-1} + \textstyle\sum_{i=1}^{3} F(S_i))\, dP$$

and again by $M \| h \|^4 n^{-3}$ by arguing as in (2.39)–(2.41). For example,

$$\int_{[J_1=J_2=3]} F(S_1)\, dP \leq \frac{\alpha(m)}{n-2} \int_{[J_1=3]} F(S_1)\, dP = \alpha(m)[n(n-1)(n-2)]^{-1}.$$

Finally,

$$\left|\int_{[J_2=3,J_1=4]} \pi\, dP\right| \le \|h\| \int_{[J_2=3,J_1=4]} |h_1^* h_2^* h_3^*|$$

$$\le (n-3)^{-1}\|h\| \int_{[J_2=3]} |h_1^* h_2^* h_3^*|$$

$$\le [(n-3)(n-2)]^{-1}\|h\|^2 E|h_1^* h_2^*| \tag{2.46}$$

$$\le Mn^{-2}\|h\|^2(E^2|h_1| + \operatorname{cov}(|h_1^*|, |h_2^*|))$$

$$\le Mn^{-2}\|h\|^2(E^2|h_1| + \|h\|^2 n^{-1})$$

by Lemma 2.11. By our discussion and (2.43)–(2.46),

$$\left|\int_{E_{12}^c} \pi\, dP\right| \le M\|h\|^2\left(\frac{E^2|h_1|}{n^2} + E^2|h_1|F(S_1) + \|h\|^2 n^{-3}\right). \tag{2.47}$$

Now by the Schwartz inequality,

$$E^2|h_1|F(S_1) \le E|h_1|E|h_1|F^2(S_1) \le \frac{E^2|h_1|}{n^2} + n^2E^2|h_1|F^2(S_1).$$

The lemma, therefore, follows from (2.31) and (2.47).

LEMMA 2.48. *For $M_5 < \infty$*

$$|E[h_1^*]^2 h_2^* h_3^*| \le M_5\|h\|^2\left(\frac{E^2|h_1|}{n} + nE^2|h_1|F(S_1) + \frac{\|h\|^2}{n^2}\right). \tag{2.49}$$

PROOF. The argument goes much as for Lemma 2.22 and is sketched. If we denote the integrand by π^*

$$\left|\int_{[J_1\neq 2,3]} \pi^*\, dP\right| \le M\|h\|\left\{(n^{-1}E|h_1| + E|h_1|F(S_1)) \times \int_{[J_1\neq 2,3]} [h_1^*]^2 + \|h\|^3 n^{-2}\right\}$$

$$\le M\|h\|^2\{n^{-1}E^2|h_1| + nE^2|h_1|F(S_1) + \|h\|^2 n^{-2}\},$$

while

$$\left|\int_{[J_1=2]} [h_1^*]^2 h_2^* h_3^*\, dP\right| \le \|h\|^2 \int_{[J_1=2]} |h_1^* h_3^*|\, dP \le Mn^{-1}\|h\|^2 \int |h_1^* h_2^*|\, dP$$

$$\le M\|h\|^2 n^{-1}(E^2|h_1| + n^{-2}\|h\|^2)$$

arguing as in (2.46). The lemma follows.

PROOF OF THEOREM. Write

$$\begin{aligned} E\left(\textstyle\sum_i h_i^*\right)^4 &\le nE[h_1^*]^4 + 6n(n-1)E[h_1^*]^2[h_2^*]^2 \\ &\quad + 6n(n-1)(n-2)|E[h_1^*]^2 h_2^* h_3^*| + n(n-1)(n-2)(n-3)|Eh_1^* h_2^* h_3^* h_4^*|. \end{aligned} \tag{2.50}$$

We apply Lemmas 2.22 and 2.48 to the last two terms of (2.50); note that the second term is

$$\le 6n^2\|h\|^2(E^2|h_1^*| + |\operatorname{cov}(|h_1^*|, |h_2^*|)|)$$

and apply Lemma 2.11, and bound $E[h_1^*]^4$ by $16\|h\|^4$. The theorem follows.

3. Second moment convergence. The central result of this section is the evaluation of the limit of $\mathrm{Var}((1/\sqrt{n})\sum_1^n h(X_j, D_j))^2$ for a certain class of functions h. Starting with the density $f(x)$, define

$$\gamma(x) = f(x)^{-1/m},$$

and for any measurable function h on $E^{(m)} \times [0, \infty) \to E_1$, let

$$\tilde{h}(x, r) = h(x, \gamma(x)r).$$

Define L_0, L_1, L_2 as functions of bounded variation given by

$$L_0(r) = e^{-V(r)} \tag{3.1}$$

$$L_1(r_1, r_2) = e^{-V(r_1)-V(r_2)}[V(r_1) + V(r_2) - V(r_1)V(r_2)] \tag{3.2}$$

$$L_2(r_1, r_2) = e^{-V(r_1)-V(r_2)}\left[\int_{B(r_1,r_2)} (e^{V(r_1,r_2,z)} - 1)\, dz - V(\max(r_1, r_2))\right] \tag{3.3}$$

where

$$B(r_1, r_2) = \{z; \max(r_1, r_2) \le \|z\| \le r_1 + r_2\}$$

$$V(r_1, r_2, z) = \int_{S(0,r_1)\cap S(z,r_2)} dy.$$

For any two functions h, h' define the functional $L(h, h')$ by

$$\begin{aligned} L(h, h') = &\int \tilde{h}(x_1, r_1)\tilde{h}'(x_2, r_2) f(x_1) f(x_2) L_1(dr_1, dr_2)\, dx_1\, dx_2 \\ &+ \int \tilde{h}(x, r_1)\tilde{h}'(x, r_2) f(x) L_2(dr_1, dr_2)\, dx. \end{aligned} \tag{3.4}$$

The moment convergence result is the following.

THEOREM 3.5. *If h is measurable on $E^{(m)} \times [0, \infty) \to E^{(1)}$ and satisfies*
(i) $\|h\| < \infty$
(ii) *the set of discontinuities of h has Lebesgue measure* 0,
then

$$\mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_1^n h(X_i, D_i)\right) \to \sigma^2(h)$$

where

$$\sigma^2(h) = \int \tilde{h}^2(x, r) f(x) L_0(dr)\, dx - \left[\int\int \tilde{h}(x, r) f(x) L_0(dr)\, dx\right]^2 + L(h, h). \tag{3.6}$$

As the proof will reveal, the first two terms of (3.6) would be the limit if the R_j were independent. The $L(h, h)$ term is contributed by the local dependence of the nearest neighbor distances.

The proof of the theorem is split into two pieces. Proposition 3.7 below shows that the diagonal terms in

$$\frac{1}{n}\left(\sum_1^n h^*(X_i, D_i)\right)^2$$

converge to the first two terms of (3.6). Then Proposition 3.20 gives convergence of the off-diagonal terms to $L(h, h)$. We assume throughout that the conditions of the theorem hold.

Let X, D be a random m vector and nonnegative random variable respectively such that

X had density f and

$$P[D > r \mid X] = \exp\{-f(X)V(r)\}.$$

Equivalently, $D/\gamma(X)$ is independent of X and

$$P[D/\gamma(X) > r] = L_0(r).$$

PROPOSITION 3.7. *Let f satisfy* A(i)–(iii). *Then, as $n \to \infty$,*

$$(X_1, D_{1n}) \to_{\mathscr{D}} (X, D)$$

where (X_1, D_{1n}) is used to stand generically for the common law of any of the pairs (X_i, D_i) and $\to_{\mathscr{D}}$ denotes convergence in distribution. Therefore

$$Eh(X_1, D_{1n}) \to \int \tilde{h}(x, r)f(x)L_0(dr)\, dx \tag{3.8}$$

$$\operatorname{Var} h(X_1, D_{1n}) \to \int \tilde{h}^2(x, r)f(x)L_0(dr)\, dx - \left(\int \tilde{h}(x, r)f(x)L_0(dr)\, dx\right)^2. \tag{3.9}$$

PROOF. Almost immediate, since

$$P(D_{1n} > r \mid X_1 = x) \to e^{-f(x)V(r)} = P(D > r \mid X = x)$$

and the set of discontinuities of h has probability zero with respect to the (X, D) distribution.

PROPOSITION 3.10. *For $h(x, r)$ any function satisfying the hypothesis of Theorem* 3.5

$$n \operatorname{Cov}(h(X_1, D_1), h(X_2, D_2)) \to L(h, h).$$

PROOF. It is, we assert, sufficient to show for any two functions ϕ_1, ϕ_2 of the form

$$\phi_i(x, r) = g_i(x)I(r \geq r_i), \qquad i = 1, 2 \tag{3.11}$$

with $g_i(x)$ uniformly continuous and bounded, that

$$n \operatorname{Cov}(\phi_1(X_1, D_1), \phi_2(X_2, D_2)) \to L(\phi_1, \phi_2). \tag{3.12}$$

To see this note that if $\mathscr{F}$ is the set of all finite linear combinations of functions of the form (3.11) then we can get a sequence $h_k \in \mathscr{F}$ such that

$$\| h_k \| \leq 2 \| h \|$$

and with respect to L-measure on $E^{(m)} \times [0, \infty)$, $h_k \to h$ a.e. (since h is a.e. continuous). Now

$$\begin{aligned} &\operatorname{Cov}(h(X_1, D_1), h(X_2, D_2)) - \operatorname{Cov}(h_k(X_1, D_1), h_k(X_2, D_2)) \\ &\qquad = \operatorname{Cov}(h(X_1, D_1) - h_k(X_1, D_1), h(X_2, D_2) + h_k(X_2, D_2)). \end{aligned} \tag{3.13}$$

Using Corollary 2.15 on (3.13) gives the bound

$$\begin{aligned} &\limsup_n | \operatorname{Cov}(h(X_1, D_1), h(X_2, D_2)) - \operatorname{Cov}(h_k(X_1, D_1), h_k(X_2, D_2)) | \\ &\qquad \leq c \| h \| (E | h - h_k |^2)^{1/2}. \end{aligned}$$

Now the bounded convergence theorem gives $E(h - h_k)^2 \to 0$, and (3.12) implies that

$$\operatorname{Cov}(h_k(X_1, D_1), h_k(X_2, D_2)) \to L(h_k, h_k).$$

Since $L(h_k, h_k) \to L(h, h)$, the assertion follows.

PROOF OF (3.12). For $i = 1, 2$, let

$$S_i = S(x_i, n^{-1/m}r_i), \qquad F_i = F(S_i), \qquad F_{12} = F(S_1 \cap S_2)$$

and let

$$A = \{(x_1, x_2); f3 x_1 - x_2\| \geq n^{-1/m}(r_1 + r_2)\}$$
$$B = \{(x_1, x_2); n^{-1/m}\max(r_1, r_2) \leq \|x_1 - x_2\| \leq n^{-1/m}(r_1 + r_2)\}$$
$$C = \{(x_1, x_2), \|x_1 - x_2\| \leq n^{-1/m}\max(r_1, r_2)\}.$$

Then

$$P(R_1 \geq n^{-1/m}r_1, R_2 \geq n^{-1/m}r_2 \mid X_1 = x_1, X_2 = x_2) = \begin{cases} (1 - F_1 - F_2)^{n-2}, & (x_1, x_2) \in A \\ (1 - F_1 - F_2 + F_{12})^{n-2}, & (x_1, x_2) \in B \\ 0, & (x_1, x_2) \in C \end{cases}$$

and

$$P(R_i \geq n^{-1/m}r_i \mid X_i = x_i) = (1 - F_i)^{n-1}.$$

Then, denoting

$$L(x_1, x_2, r_1, r_2) = P(R_1 \geq n^{-1/m}r_1, R_2 \geq n^{-1/m}r_2 \mid X_1 = x_1, X_2 = x_2) - [(1 - F_1)(1 - F_2)]^{n-1}$$

and $g_i(x_i)$ by g_i, $f(x_i)$ by f_i,

$$\begin{aligned} \operatorname{Cov}(\phi_1, \phi_2) &= \int g_1(x_1)g_2(x_2)L(x_1, x_2, r_1, r_2)f(x_1)f(x_2)\, dx_1 dx_2 \\ &= \int g_1g_2[(1 - F_1 - F_2)^{n-2} - (1 - F_1)^{n-1}(1 - F_2)^{n-1}]f_1f_2 \\ &\quad + \int_B g_1g_2[(1 - F_1 - F_2 + F_{12})^{n-2} - (1 - F_1 - F_2)^{n-2}]f_1f_2 \\ &\quad - \int_C g_1g_2[(1 - F_1 - F_2)^{n-2}]f_1f_2 \\ &= I_1 + I_2 - I_3. \end{aligned}$$

Because $nF_i \leq \bar{f}V(r_i)$, where $\bar{f}$ is the supremum of f, and $nF_i \to f(x_i)V(r_i)$, for fixed x_1, x_2

$$\begin{aligned} &n[(1 - F_1 - F_2)^{n-2} - (1 - F_1)^{n-1}(1 - F_2)^{n-1}] \\ &\quad = n(1 - F_1)^{n-2}(1 - F_2)^{n-2}\left[\left\{1 - \frac{F_1F_2}{(1 - F_1)(1 - F_2)}\right\}^{n-2} - (1 - F_1)(1 - F_2)\right] \\ &\quad \to e^{-f(x_1)V(r_1)-f(x_2).V(r_2)}[f(x_1)V(r_1) + f(x_2)V(r_2) - f(x_1)f(x_2)V(r_1)V(r_2)]. \end{aligned}$$

Furthermore, the convergence is bounded. Therefore

$$nI_1 \to \int \tilde{\phi}(x_1, r_1)\tilde{\phi}(x_2, r_2)L_1(dr_1, dr_2)f(x_1)f(x_2)\, dx_1\, dx_2$$

as can be seen by making the transformations $V(r_i') = f(x_i)V(r_i)$.

In I_2, I_3 make the transformation

$$x_2 = x_1 + n^{-1/m} z,$$

leading to

$$B = \{(x_1, z); \max(r_1, r_2) \le \|z\| \le r_1 + r_2\}$$
$$C = \{(x_1, z); \|z\| \le \max(r_1, r_2)\}.$$

On BUC, for x_1 fixed

$$f(x_2)g_2(x_2) \to f(x_1)g_2(x_1)$$

uniformly, and

$$nF_i \to f(x_1)V(r_1), \qquad nF_{12} \to f(x_1)V(r_1, r_2, z)$$

where

$$V(r_1, r_2, z) = \int_{[\|y\| \le r_1, \|y-z\| \le r_2} dy.$$

Therefore

$$nI_2 \to \int \left[\int_B (e^{f(x) V(r_1, r_2, z)} - 1)\, dz \right] e^{-f(x)[V(r_1) + V(r_2)]} g_1(x) g_2(x) f^2(x)\, dx.$$

A simpler argument gives

$$nI_3 \to \int V(\max(r_1, r_2)) e^{-f(x)[V(r_1) + V(r_2)]} g_1(x) g_2(x) f^2(x)\, dx.$$

In both integrals, make the substitution $V(r_i') = f(x)V(r_i)$ and add the limits together to get the proposition.

4. A central limit theorem. The main result of this section is

THEOREM 4.1. *Suppose the set of discontinuities of h has Lebesgue measure* 0 *in* $E^{(m)} \times [0, \infty)$ *and*

$$\sup_{x,d}|h| = \|h\| < \infty.$$

Then if the density of the distribution satisfies A(i)–(iii),

$$\frac{1}{\sqrt{n}} \sum_1^n h^*(X_j, D_j) \to_{\mathscr{D}} N(0, \sigma^2(h)) \tag{4.2}$$

where $\sigma^2(h)$ is given in Theorem 3.5.

The proof proceeds in a series of propositions.

NOTATIONAL CONVENTION. Lower case c denotes a constant depending only on m and $\|h\|$. The dependence of other constants on various auxiliary parameters introduced below will be noted as needed.

PROPOSITION 4.3. *There exists a sequence of bounded sets $C_N \subset E^{(m)}$ with $C_N \subset C_{N+1}$ such that*

1) $\text{diameter}(C_N) \le N$
2) $\inf_{x \in C_N} f(x) = \delta_N > 0$
3) $P(X \in C_N^c) \to 0$.

PROOF. There exist compact sets $A_N \subset A_{N+1}$ such that $\int_{A_N} f\, dx \to 1$. Choose $\delta_N > 0$

such that $\delta_N \int_{A_N} dx \to 0$. Let

$$F_N = \{x; f(x) \geq \delta_N\}$$

and take $C_N = A_N \cap F_N$. Then

$$\int_{A_N} f - \int_{C_N} f \leq \int_{A_N \cap F_N^c} f \leq \delta_N \int_{A_N} dx$$

so $\int_{C_N} f \to 1$.

In preparation for the next step, let D_N be a cube of side N such that $C_N \subset D_N$. Divide D_N into $L = (k)^m$ congruent subcubes $D_{N,\ell}$, $\ell = 1, \cdots, L$, and let

$$B_\ell = \bar{D}_{N,\ell} \cap C_N, \qquad \ell = 1, \cdots, L$$

$$\tilde{B} = \cup_\ell \partial(B_\ell)$$

where ∂ denotes boundary. The B_ℓ, $\ell = 1, \cdots, L$ provide the basic cells such that nearest neighbor links between different cells will be cut. *From now on until the end of the string of propositions N and the B_ℓ, $\ell = 1, \cdots, L$ will be fixed.*

Select $d_N > 0$ and let

$$E_N = \{x; x \in C_N, d(x, \tilde{B}) \geq d_N\}$$

where $d(x, \tilde{B})$ is the distance from x to the set $\tilde{B}$. Write (X, D) for (X_1, D_{1n}). Note that by using $f(x) \leq \sup_x f(x) = \bar{f}$, we get

$$P(X \in C_N, d(X, \tilde{B}) < d_N) \leq 2md_N L^{1/m} N^{m-1} \bar{f}.$$

Now let

$$\mathbf{h}(x, d) = I(x \in E_n) h(x, d).$$

We suppress dependence on N, L here and in the sequel except where emphasis is needed. Denote (recalling that $h^* = h - Eh$, $\mathbf{h}^* = \mathbf{h} - E\mathbf{h}$),

$$Z_n = \frac{1}{\sqrt{n}} \sum_1^n h^*(X_j, D_j), \qquad Z_n(N, L) = \frac{1}{\sqrt{n}} \sum_1^n \mathbf{h}^*(X_j, D_j).$$

PROPOSITION 4.4. $E(Z_n - Z_n(N, L))^2 \leq c(P(X \in E_N^c))^{1/2}$.

PROOF. This follows directly from Corollary 2.15.

For the next step define

$$R_j' = \begin{cases} 0 & \text{if } X_j \in B_\ell, \text{ no other } X_i \in B_\ell \\ \inf_{i \neq j, X_i \in B_\ell} \|X_i - X_j\| & \text{if } X_j \in B_\ell \end{cases}$$

and redefine $h(x, 0) = 0$. Let $D_j' = n^{1/m} R_j'$ and

$$Z_n'(N, L) = \frac{1}{\sqrt{n}} \sum_1^n \mathbf{h}^*(X_j, D_j').$$

PROPOSITION 4.5. $E(Z_n(N, L) - Z_n'(N, L))^2 \leq cne^{-(n-1)\varepsilon_N V(d_N)}$ *where* $\varepsilon_N > 0$ *depends only on* N.

PROOF.

$$E(Z_n(N, L) - Z_n'(N, L))^2 \leq \frac{1}{n} E(\textstyle\sum_j \Delta_j)^2 \leq \sum_j E\Delta_j^2$$

where

$$\Delta_j = \mathbf{h}(X_j, D_j) - \mathbf{h}(X_j, D_j') - E(\mathbf{h}(X_j, D_j) - \mathbf{h}(X_j, D_j'))$$

so

$$E(Z_n(N, L) - Z'_n(N, L))^2 \le \sum_j E(\mathbf{h}(X_j, D_j) - \mathbf{h}(X_j, D'_j))^2.$$

Now $X_j \in E_N$ and $d(X_j, \tilde{B}) > R_j$ implies $R'_j = R_j$. So

$$\begin{aligned} E(Z_n(N, L) - Z'_n(N, L))^2 &\le 2\|h\|^2 \sum_j P(R_j \ne R'_j, X_j \in E_N) \\ &\le 2\|h\|^2 nP(d(X, \tilde{B}) \le R, X \in E_N) \end{aligned}$$

where (X, R) stands for (X_1, R_{1n}) by our usual convention. Now

$$P(R \ge r \mid X = x) = [1 - F(S(x, r))]^{n-1}.$$

Note that $d(X, \tilde{B}) \le N\sqrt{m}$ for $X \in E_N$. Now

$$\inf_{x \in C_N} \inf_{0 \le r \le \sqrt{m}N} [F(S(x, r))/V(r)] = \varepsilon_N > 0$$

since $M(r, x) = F(S(x, r))/V(r)$ is jointly continuous on $[0, \sqrt{m}\ N] \times \bar{C}_N$, where $\bar{C}_N$ is the closure of C_N, and since $M(r, x) > 0$ everywhere in $\bar{C}_N \times [0, \sqrt{m}\ \ N]$. Therefore

$$P(R \ge d(X, \tilde{B}), X \in E_N) \le \int_{X \in E_N} e^{-(n-1)\varepsilon_N V(d(x, \tilde{B}))} f(x)\, dx.$$

For $x \in E_N$, $d(x, \tilde{B}) \ge d_N$, so

$$P(R \ge d(X, \tilde{B}), X \in E_N) \le e^{-(n-1)\varepsilon_N V(d_N)}$$

and the proposition follows.

For the next step, put $B_0 = C_N^c$, and denote

$$P(X \in B_\ell) = p_\ell, \qquad \ell = 0, 1, \cdots, L$$

so $\sum_{\ell=1}^{L} p_\ell = 1$. (Assume that for every $\ell, p_\ell > 0$, otherwise delete B_ℓ.) Let

$$n_\ell = \#(X_j \in B_\ell)$$

so the $(n_0, \cdots, n_L)$ have a multinomial distribution with parameters$(p_0, \cdots, p_L)$. Consider the following construction: draw numbers $n_0, \cdots, n_L$, $\sum n_\ell = n$ from a multinomial distribution with parameters $(p_0, \cdots, p_L)$. Then put n_ℓ points $X_i^{(\ell)}$, $i = 1, \cdots, n_\ell$ into B_ℓ using the distribution

$$F_\ell(dx) = P(X \in dx \mid X \in B_\ell).$$

Denote by P_ℓ the joint distribution of $X_i^{(\ell)}$, $i = 1, \cdots, n_\ell$, let $R_i^{(\ell)}$ be the nearest neighbor distance to $X_i^{(\ell)}$ from the other points in B_ℓ, and $D_i^{(\ell)} = n^{1/m} R_i^{(\ell)}$. Put

$$T_\ell = \begin{cases} \sum_{i=1}^{n_\ell} \mathbf{h}(X_i^{(\ell)}, D_i^{(\ell)}), & n_\ell > 1 \\ 0, & n_\ell \le 1. \end{cases}$$

Then

$$\sum_{\ell=1}^{L} T_\ell = \sum_{j=1}^{n} \mathbf{h}(X_j, D'_j).$$

PROPOSITION 4.6. *There are constants $\gamma_{n,\ell}$, $\ell = 1, \cdots, L$ such that $\gamma_{n,\ell} \to \gamma_\ell$ and*

$$E(E(T_\ell \mid n_\ell) - ET_\ell - (n_\ell - En_\ell)\gamma_{n,\ell})^2 \le C(\ell) < \infty$$

where $C(\ell)$ is independent of n.

PROOF. Define

$$W_\ell(r \mid x, n_\ell) = P_\ell(n^{1/m} R_1^{(\ell)} > r \mid X_1^{(\ell)} = x) = [1 - F_\ell(S(x, rn^{-1/m}))]^{n_\ell - 1}.$$

Note that

$$E(T_\ell \mid n_\ell) = n_\ell \int \mathbf{h}(x, r) W_\ell(dr \mid x, n_\ell) F_\ell(dx).$$

Define

$$\chi_n(r \mid x) = W_\ell(r \mid x, np_\ell) = [1 - F_\ell(S(x, rn^{-1/m}))]^{np_\ell - 1}$$

and suppressing the dependence on L, let

$$\mu_n = (n_\ell - np_\ell)/(np_\ell - 1).$$

Then

$$W_\ell(r \mid x, n_\ell) = \chi_n^{\mu_n + 1}.$$

Then

$$W_\ell(dr \mid x_1 n_\ell) = \frac{n_\ell - 1}{np_\ell - 1} \chi_n^{\mu_n} \chi_n(dr \mid x) = (\mu_n + 1)\chi_n^{\mu_n}\, d\chi_n$$

where $d\chi_n \equiv \chi_n(dr \mid x)$. This is zero for $\mu_n = -1$, so we eliminate this set in the expectations to follow. Writing $n_\ell = (np_\ell - 1)\,\mu_\ell + np_\ell$ leads to the expression

$$E(T_\ell \mid n_\ell) = np_\ell(1 + \mu_n)^2 \int \mathbf{h}\chi_n^{\mu_n}\, d\chi_n\, dP_\ell - \mu_n(1 + \mu_n) \int \mathbf{h}\chi_n^{\mu_n}\, d\mu_n\, dP_\ell. \tag{4.7}$$

The expectation of the square of the second term in (4.7) above is bounded by $C_\ell \|\mathbf{h}\|^2/n$, and is henceforth ignored.

Next, expand

$$\chi_n^{\mu_n} = 1 + \mu_n \log \chi_n + \frac{\mu_n^2}{2} (\log \chi_n)^2 \chi_n^{\theta\mu_n},$$

where $0 \le \theta \le 1$, and substitute into the first term of (4.7). We assert that all terms containing a power of μ_n higher than one have squares whose expectations are uniformly bounded in n. For example

$$(np_\ell)^2 E\left(\mu_n^2 \int \mathbf{h}(\log \chi_n)\, d\chi_n\, dP_\ell\right)^2 \le (np_\ell)^2 \|h\| E\mu_n^4 \le C\|h_1\|^2(1 - p_\ell)^2$$

and

$$\begin{aligned} &(np_\ell)^2 E\left(\mu_n^2(1 + \mu_n)^2 \int \mathbf{h}(\log \chi_n)^2 \chi_n^{\theta\mu_n}\, d\chi_n\, dP_\ell\right)^2 \\ &\quad \le \|\mathbf{h}\|^2 (np_\ell)^2 E\left(\mu_n^2(1 + \mu_n)^2 \int (\log \chi_n)^2 \chi_n^{\theta\mu_n}\, d\chi_n\, dP_\ell\right)^2 \\ &\quad \le 2\,\|\mathbf{h}\|^2 (np_\ell)^2 [E\{\mu_n^4(1 + \mu_n)^{-2}; -1 < \mu_n \le 0\} + E\{\mu_n^4(1 + \mu_n^4); \mu_n > 0\}] \\ &\quad \le C_\ell \|\mathbf{h}\|^2. \end{aligned}$$

Therefore

$$E(T_\ell \mid n_\ell) = np_\ell \int \mathbf{h}(1 + \mu_n(2 + \log \chi_n))\, d\chi_n m\, dP_\ell + O_2(1) \tag{4.8}$$

so

$$E(T_\ell \mid n_\ell) - ET_\ell = np_\ell \mu_n \int \mathbf{h}(2 + \log \chi_n)\, d\chi_n\, dP_\ell + O_2(1) \tag{4.9}$$

where $O_2(1)$ in (4.8) and (4.9) denote quantities such that $\sup_n E(O_2(1))^2 < \infty$. Letting the $\gamma_{n,\ell}$ of the proposition be defined by

$$\gamma_{n,\ell} = \frac{np_\ell}{np_\ell - 1} \int \mathbf{h}(2 + \log \chi_n)\, d\chi_n\, dP_\ell.$$

The proof will be completed by showing that the integral on the right above converges.

For x fixed, $\chi_n(r\,|\,x)$ is a non-increasing function of r such that for $x \in \mathrm{Int}(B_\ell)$

$$\chi_n(r\,|\,x) \to e^{-f(x)V(r)} = \chi_0(r\,|\,x).$$

Since $\mathbf{h}(x, r)$ is a.s. continuous with respect to $d\chi_0\, dP_\ell$, then

$$\int \mathbf{h}\, d\chi_n\, dP_\ell \to \int \mathbf{h}\, d\chi_0\, dP_\ell.$$

Now let

$$\tilde{\chi}_n(r\,|\,x) = (1 - \log \chi_n(r\,|\,x))\chi_n(r\,|\,x)$$

so that

$$\tilde{\chi}_n(dr\,|\,x) = -(\log \chi_n(r\,|\,x))\chi_n(dr\,|\,x).$$

For $x \in \mathrm{Int}(B_\ell)$

$$\tilde{\chi}_n(r\,|\,x) \to (1 + f(x)V(r))e^{-f(x)V(r)} = \tilde{\chi}_0(r\,|\,x)$$

and so

$$\int \mathbf{h}(\log \chi_n)\, d\chi_n\, dP_\ell \to -\int \mathbf{h}\, d\chi_0\, dP_\ell. \tag{4.10}$$

PROPOSITION 4.11. $\frac{1}{\sqrt{n}} \sum_{\ell=1}^{L} [E(T_\ell\,|\,n_\ell) - E(T_\ell)] \to_{\mathscr{D}} N(0, \sigma_{N,L}^2)$ *where*

$$\sigma_{N,L}^2 = \sum_\ell \gamma_\ell^2 p_\ell - \left(\sum \gamma_\ell p_\ell\right)^2.$$

Moreover, $n^{-1}(\sum_{\ell=1}^{L} [E(T_\ell\,|\,n_\ell) - E(T_\ell)]^2) \to \sigma_{N,L}^2$.

PROOF. Clear from the preceding proposition.

It is useful to recall the dependence of parameters on N and L at this point.

PROPOSITION 4.12. *Let*

$$U_n = \frac{1}{\sqrt{n}} \sum_{\ell=1}^{L} (T_\ell - E(T_\ell\,|\,n_\ell)). \tag{4.13}$$

Then there is a constant $s_{N,L}^2 < \infty$ *such that*

$$E(U_n^2\,|\,n_1, \cdots, n_L) \to_{\text{a.s.}}^{L_1} s_{N,L}^2.$$

PROOF. Given $\mathbf{n} = n_1, \cdots, n_L$, the terms in the sum for U_n are independent. Thus

$$E(U_n^2\,|\,n_1, \cdots, n_L) = \frac{1}{n} \sum_\ell \mathrm{Var}(T_\ell\,|\,n_\ell),$$

and

$$\mathrm{Var}(T_\ell\,|\,n_\ell) = n_\ell \mathrm{Var}(\mathbf{h}(X_1^{(\ell)}, D_1^{(\ell)})\,|\,n_\ell) + n_\ell(n_\ell - 1)\mathrm{Cov}(\mathbf{h}(X_1^{(\ell)}, D_1^{(\ell)}), \mathbf{h}(X_2^{(\ell)}, D_2^{(\ell)})\,|\,n_\ell);$$

it is then sufficient to show that

$$\mathrm{Var}(\mathbf{h}(X_1^{(\ell)}, D_1^{(\ell)})\,|\,n_\ell) \to_{\text{a.s.}}^{L_1} \text{constant}$$

$$n\,\mathrm{Cov}(\mathbf{h}(X_1^{(\ell)}, D_1^{(\ell)}), \mathbf{h}(X_2^{(\ell)}, D_2^{(\ell)})\,|\,n_\ell) \to_{\text{a.s.}}^{L_1} \text{constant}.$$

This result can be gotten through a simple modification of Propositions 3.7 and 3.10.

Now we are ready for the final steps. We can write

$$Z_n'(N, L) =_{\mathscr{D}} U_n + V_n, \tag{4.14}$$

with U_n defined in (4.13) and

$$V_n = \frac{1}{\sqrt{n}} \sum_{\ell=1}^{L} [E(T_\ell \mid n_\ell) - ET_\ell].$$

By $=_{\mathscr{D}}$ we mean equality in distribution when U_n and V_n have the joint distribution we have implicitly given them. Denote $e_N^2 = P(X \in E_N^c)$.

PROPOSITION 4.15. *If $\sigma^2 = \lim_n \mathrm{Var}(Z_n)$, then*

$$|\sigma^2 - (s_{N,L}^2 + \sigma_{N,L}^2)| \le ce_N + 2\sigma\sqrt{ce_N}.$$

PROOF. By Propositions 4.4 and 4.5

$$\limsup_n E(Z_n - Z_n'(N, L))^2 \le ce_N. \tag{4.16}$$

Use the inequality

$$|EZ_n^2 - EZ_n'^2(N, L)| \le E|Z_n - Z_n'(N, L)|^2 + 2\sqrt{E(Z_n)^2 E(Z_n - Z_n'(N, L))^2} \tag{4.17}$$

and take $n \to \infty$ to get the result.

PROPOSITION 4.18. *Let $\alpha = \sqrt{\max_\ell p_\ell}$ and take $|t|^3 \le \alpha^{-1}$. Note that α depends on both N and L. Let $g_n(t; N, L)$ denote the characteristic function of $Z_n'(N, L)$. Then*

$$\limsup_n |g_n(t; N, L) - e^{-(\sigma_{N,L}^2 + s_{N,L}^2)t^2/2}| \le c\alpha|t|^3.$$

PROOF.

$$g_n(t; N, L) = Ee^{it(U_n + V_n)} = E(e^{itV_n} E(e^{itU_n} \mid \mathbf{n})), \qquad \mathbf{n} = (n_0, \cdots, n_L).$$

Given $\mathbf{n}$, $U_n = \sum_1^L A_\ell$, with the A_ℓ independent and having the conditional distribution of $T_\ell - E(T_\ell \mid n_\ell)$ given n_ℓ. Hence

$$E(e^{itU_n} \mid \mathbf{n}) = \Pi f_\ell(t), \qquad f_\ell(t) = E(e^{itA_\ell} \mid n_\ell).$$

Applying Corollary 2.3 to A_ℓ,

$$E(A_\ell^2 \mid n_\ell) \le c_1(n_\ell/n), \qquad E(|A_\ell^3| \mid n_\ell) \le c_2(n_\ell/n)^{3/2}$$

where c_k will denote constants depending only on m, $\|h\|$, and θ_k will be quantities such that $|\theta_k| \le 1$. Then

$$|1 - f_\ell(t)| \le \frac{t^2}{2} E(A_\ell^2 \mid n_\ell) \le (c_1/2)t^2(n_\ell/n)$$

$$|f_\ell(t) - 1 + \frac{t^2}{2} E(A_l^2 \mid n_\ell)| \le c_2|t|^3(n_\ell/n)^{3/2}.$$

Temporarily restrict t to the range $|t|\alpha \le c_1^{-1/2}/2$. Define

$$B_n = \{\max_\ell(n_\ell/n) \le 2 \max_\ell p_\ell\}.$$

On B_n, $|1 - f_\ell(t)| \le 1/4$, hence

$$\log f_\ell(t) = \log[1 - (1 - f_\ell(t))] = -\frac{t^2}{2} E(A_\ell^2 \mid n_\ell) + \theta_1 c_2|t^3|(n_\ell/n)^{3/2} + \theta_2 c_3 t^4(n_\ell/n)^2.$$

So

$$\Pi f_\ell(t) = \exp\left(-\frac{t^2}{2} \sum_\ell E(A_\ell^2 \mid n_\ell) + \Delta_n\right)$$

where, since $|t^3|\alpha \le 1$

$$|\Delta_n| \le c_2|t^3| \sum_\ell (n_\ell/n)^{3/2} + c_3 t^4 \sum (n_\ell/n)^2 \le c_2|t^3|\alpha + c_3|t^4|\alpha^2 \le c_4|t|^3\alpha.$$

Therefore

$$|e^{\Delta_n} - 1| \le c_5 |t|^3 \alpha$$

and so, denoting $\beta_n^2 = E(U_n^2 | \mathbf{n})$

$$|\Pi f_\ell(t) - e^{-\beta_n^2 t^2/2}| \le c_5 |t|^3 \alpha$$

holds on B_n for all t such that $|t^3| \le \alpha^{-1}$, and $|t|\alpha \le c_1^{-1/2}/2$. Write

$$g_n(t; N, L) = E(I(B_n)e^{it(U_n+V_n)}) + E(I(B_n^c)e^{it(U_n+V_n)}).$$

Since $P(B_n^c) \to 0$, the second term goes to zero, so

$$\limsup |g_n(t; N, L) - Ee^{itV_n - \beta_n^2 t^2/2}| \le c_5 |t^3| \alpha.$$

Combining this with Propositions 4.11 and 4.12

$$\limsup |g_n(t; N, L) - e^{-(s_{N,L}^2 + \sigma_{N,L}^2)t^2/2}| \le c_5 |t^3| \alpha.$$

To complete the proof we need only remove the restriction $|t|\alpha \le c_1^{-1/2}/2$. But this can clearly be done by increasing the constant c_5.

The stage is now set for the proof of Theorem 4.1. By (4.16)

$$\limsup_n |g_n(t) - g_n(t; N, L)| \le \limsup_n E|\exp\{it(Z_n - Z_n'(N, L))\} - 1| \le |t| \sqrt{ce_N},$$

where $g_n(t)$ is the characteristic function of Z_n. So, by Proposition 4.18,

$$\limsup_n \left| g_n(t) - \exp\left\{ -(s_{N,L}^2 + \sigma_{N,L}^2)\frac{t^2}{2} \right\} \right| \le c(|t|^3 \alpha + |t| \sqrt{e_N}) \tag{4.19}$$

for $|t|^3\alpha \le 1$. Now let $N \to \infty$, $L \to \infty$ in such a way that $\alpha \to 0$ and $e_N \to 0$. By Proposition 4.15, if $e_N \to 0$, uniformly in L,

$$\lim_N (s_{N,L}^2 + \sigma_{N,L}^2) = \sigma^2.$$

Since the restriction $|t|^3\alpha \le 1$ is satisfied eventually for any fixed t, as $\alpha \to 0$ we conclude that, for all t,

$$\lim_n g_n(t) = e^{-\sigma^2 t^2/2}$$

and (4.1) follows since the equality of σ^2 and $\sigma^2(h)$ is derived from the moment convergence theorem 3.5.

By considering linear combinations of h's it is clear how the results can be generalized to provide a multidimensional central limit theorem, and the moment convergence theorem 3.5 can be easily modified to give the limiting form of the covariance matrix.

5. The process $\hat{H}(t)$ and goodness-of-fit. First, a Glivenko-Cantelli type theorem is established for $H(t)$. Let

$$\lambda(x) = \begin{cases} \dfrac{f(x)}{g(x)}; & g(x) > 0 \\ \infty; & g(x) = 0 \end{cases} \tag{5.1}$$

and define a d.f. H by,

$$H(t) = \begin{cases} Et^{\lambda(X_1)}, & 0 \le t < 1 \\ 1, & t \ge 1, \end{cases} \tag{5.2}$$

and

$$\alpha = H(1) - H(1-) = P[g(X_1) = 0]. \tag{5.3}$$

Note that if $f = g$, then $\alpha = 0$ and H is the d.f. of the uniform distribution.

THEOREM 5.4. *If* A(iii) *holds, as* $n \to \infty$,

$$\sup_y |\hat{H}(y) - H(y)| \to_{\text{a.s.}} 0. \tag{5.5}$$

PROOF. We begin by showing,

$$\hat{H}(y) \to H(y) \quad \text{a.s.} \quad \forall 0 \le y < 1 \tag{5.6}$$

and

$$\hat{H}(1-) \to 1 - \alpha = H(1-), \quad \text{a.s.} \tag{5.7}$$

To prove (5.6) note that by Corollary 2.3,

$$P[|\hat{H}(y) - E\hat{H}(y)| \ge \varepsilon] = O(n^{-2})$$

and hence by the Borel-Cantelli lemma,

$$\hat{H}(y) - E\hat{H}(y) \to 0 \quad \text{a.s.} \quad \forall 0 \le y < 1. \tag{5.8}$$

Assertion (5.6) then follows by using (3.7) to show that $E\hat{H}(y) \to H(y)$. Next (5.7) is an immediate consequence of the S.L.L.N. To complete the proof of the theorem, let

$$\hat{H}^*(y) = \begin{cases} \dfrac{\hat{H}(y)}{\hat{H}(1-)}, & 0 \le y < 1 \\ 1, & y \ge 1 \end{cases} \tag{5.9}$$

and define H^* similarly in relation to H. By (5.6) and (5.7) $\hat{H}^*$ converges in law to H^* with probability 1. But H^* is continuous and hence by Polya's theorem,

$$\sup_y |\hat{H}^*(y) - H^*(y)| \to_{\text{a.s.}} 0 \tag{5.10}$$

and (5.5) follows from (5.10) and (5.7).

Define a stochastic process on [0, 1] by,

$$Z_n(t) = \sqrt{n}(\hat{H}(t) - E\hat{H}(t)), \qquad 0 \le t \le 1, \tag{5.11}$$

and a corresponding Gaussian process Z with mean 0 whose covariance function $\gamma(s, t)$, $s \le t$, is defined by

$$\begin{aligned} \gamma(s, t) = &\int fs^{\lambda}\left(1 - \int ft^{\lambda}\right) \\ &- \left(\log s \int \lambda s^{\lambda} f \int t^{\lambda} f + \log t \int \lambda t^{\lambda} f \int s^{\lambda} f + \log s \log t \int t^{\lambda} f \int s^{\lambda} f\right) \\ &+ \log s \int \lambda (st)^{\lambda} f + \int \lambda (st)^{\lambda} f \int_{B(s,t)} (\eta^{\lambda}(s, t, w) - 1)\, dw\, dx \end{aligned} \tag{5.12}$$

(We write λ, f for $\lambda(x)$, $f(x)$ etc.)
where

$$B(s, t) = \{w : r_1 \le \|w\| \le r_1 + r_2\}; \qquad \log \eta(s, t, w) = \int_{S(0,r_1) \cap S(w,r_2)} dz$$

where

$$V(r_1) = -\log s; \qquad V(r_2) = -\log t.$$

If $f = g$, then $\gamma(s, t)$, $s \le t$, reduces to

$$\gamma(s, t) = s - st(1 + \log t + \log s \log t) + st \int_{B(s,t)} (\eta(s, t, w) - 1)\, dw. \tag{5.13}$$

Clearly the processes $Z_n(\cdot)$ can be identified with probability measures on $D[0, 1]$ and it will follow as a consequence of our proof that $Z(\cdot)$ can be as well. In fact, if $\alpha = 0$, $Z(\cdot)$ has a.s. continuous sample functions. Our main result is

THEOREM 5.14. *Suppose that* A *and* B *hold. Then,*

$$Z_n \to Z$$

in the sense of weak convergence in $D[0, 1]$ *where* Z *is as above and has* a.s. *continuous sample functions.*

Before giving the proof we state and prove the corollary of greatest interest to us. Let

$$S_0 = n \int_0^1 (\hat{H}(t) - E\hat{H}(t))^2 \, dt$$

$$S_1 = n \int_0^1 (\hat{H}(t) - E\hat{H}(t))^2 \, d\hat{H}(t) = \sum_{j=1}^n \left| (E\hat{H})(W_{(j)}) - \frac{j}{n} \right|^2 .$$

COROLLARY 5.15. *If* $f = g$ *and* A *holds, both* S_0 *and* S_1 *tend in law to* $\int_0^1 Z^2(t)\, dt$ *where* Z *has covariance function* (5.13).

The corollary is, for S_0, an immediate consequence of Theorem 5.2. By writing

$$S_1 = \int_0^1 Z_n^2(\hat{H}^{-1}(t)) \, dt$$

we see that the corollary follows in this case from Theorems 5.1 and 5.2.

NOTES 1) The theorem can be extended to the case $\alpha > 0$ by a conditioning argument as in Section 2. Of course the Z process is then continuous only on $[0, 1)$ and has a jump at 1.

2) It is not possible in Theorem 5.1 to replace $E\hat{H}$ in the definition of Z_n by H. Although $E\hat{H}(t) \to H(t)$, the difference is of the order of $n^{-2/m}$ and will not be negligible for $m > 3$.

PROOF OF THEOREM 5.14. We begin by establishing the tightness of the Z_n sequence using the 4th moment bound proven in Section 2. Let $R_1, \cdots, R_n$ be as in Section 2 and recall that

$$D_i = n^{1/m} R_i, \qquad i = 1, \cdots, n.$$

LEMMA 5.16. *If* A(iii) *and* B *hold, the sequence of processes* $\{Z_n\}$ *is tight in* $D[0, 1]$ *and any weak limit point is in* $C[0, 1]$.

PROOF. We use a device due to Shorack (1973). Note that:

$$Z_n(t) = n^{-1/2} \sum_{i=1}^n \left(I\left(g(X_i) D_i^m < \frac{-\log t}{K_m} \right) - P\left(g(X_i) D_i^m < \frac{-\log t}{K_m} \right) \right)$$

where K_m is the volume of the unit sphere in E^m. Let

$$Q_n(t) = G_n\left(\frac{-\log t}{K_m} \right).$$

where G_n is given in Corollary 2.5. Note that by B and the dominated convergence theorem

G_n is continuous. For given $\delta > 0$, let $t_1 < \cdots < t_K$ be such that

$$Q_n(t_i) = \frac{i\delta}{\sqrt{n}}, \qquad 1 \le i \le K$$

where $\dfrac{K\delta}{\sqrt{n}} \le 1 < (K+1)\dfrac{\delta}{\sqrt{n}}$.

Let

$$Z_n^*(t) = Z_n(t_i) + \frac{\sqrt{n}}{\delta}(Q_n(t) - Q_n(t_i))(Z_n(t_{i+1}) - Z_n(t_i))$$
$$\text{for} \quad t_i \le t < t_{i+1}, \qquad 0 \le i \le K, \qquad t_0 = 0, \qquad t_{K+1} = 1.$$

Note that

$$Z_n^*(0) = Z_n^*(1) = 0.$$

An elementary application of Corollary 2.5 shows that,

$$E(Z_n^*(t) - Z_n^*(s))^4 \le M(Q_n(t) - Q_n(s))^2, \qquad \textbf{all} \quad s, t \tag{5.17}$$

where M depends on δ but is independent of n. Since, under A(iii) and B, dominated convergence implies that for each y,

$$G_n(y) \to \int f(x)\left(1 - \exp\left\{\frac{-1}{2}\frac{f(x)K_m y}{g(x)}\right\}\right) dx$$

a continuous probability distribution; it follows from a slight modification of Billingsley (1968, Theorems 12.3 and 12.4) that $\{Z_n^*\}$ is tight and that all limit points of $\{Z_n^*\}$ are in $C[0, 1]$. Next note that

$$\begin{aligned}
&\sup_t |Z_n(t) - Z_n^*(t)| \\
&\le \max\Big\{\sup\{|Z_n(t) - Z_n(t_i)| : t_i \le t < t_{i+1}\} \\
&\qquad + \frac{\sqrt{n}}{\delta}(\sup\{|Q_n(t) - Q_n(t_i)| : t_i \le t < t_{i+1}\})|Z_n(t_{i+1}) - Z_n(t_i)| : 0 \le i \le K\Big\} \qquad (5.18)\\
&\le \max\{[|Z_n(t_{i+1}) - Z_n(t_i)| + \sqrt{n}(E\hat{H}_n(t_{i+1}) - E\hat{H}_n(t_i))] \\
&\qquad + |Z_n(t_{i+1}) - Z_n(t_i)| : 0 \le i \le K\}
\end{aligned}$$

using the monotonicity of $\hat{H}_n(\cdot)$, $E\hat{H}_n(\cdot)$, $Q_n(\cdot)$. Next note that integrating (2.8) for $j = 0$, implies that for C independent of n, δ,

$$\sqrt{n}E(\hat{H}_n(t_{i+1}) - \hat{H}_n(t_i)) \le C\sqrt{n}(Q_n(t_{i+1}) - Q_n(t_i)) \le C\delta.$$

Hence,

$$\sup_t |Z_n(t) - Z_n^*(t)| \le 2\max\{|Z_n^*(t_{i+1}) - Z_n^*(t_i)| : 0 \le i \le K\} + C\delta. \tag{5.19}$$

But in view of (5.17), some elementary inequalities give

$$\begin{aligned}
&P[\max\{|Z_n^*(t_{i+1}) - Z_n^*(t_i)| : 0 \le i \le K\} \ge \varepsilon] \qquad (5.20)\\
&\qquad \le \varepsilon^{-4} M \textstyle\sum_{i=0}^K (Q_n(t_{i+1}) - Q_n(t_i))^2 \le M\dfrac{\delta}{\sqrt{n}} \to 0.
\end{aligned}$$

By (5.18)–(5.20) for each $\delta > 0$, C independent of δ

$$P[\sup_t |Z_n(t) - Z_n^*(t)| > 2C\delta] \to 0. \tag{5.21}$$

Since $\{Z_n^*\}$ is tight for each δ, (5.21) implies tightness of $\{Z_n\}$ and a.s. continuity of all limit points. (See, for example, Theorem 4.2 of Billingsley (1968). Note that the dependence of Z_n^* on δ is immaterial.)

Asymptotic normality of $(Z_n(t_1), \cdots, Z_n(t_n))$ follows from the representation given in the introduction,

$$Z_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h^*(X_i, D_i)$$

with

$$h(x, d) = I(\exp\{-g(x)V(d)\} \le t)$$

and the multivariate extension of Theorem 4.1. Similarly the formulae (5.11) and (5.12) for $\gamma(s, t)$ may be obtained after tedious calculations from the appropriate straightforward generalizations of Proposition 3.10.

As an immediate consequence of Theorem 5.4 and Corollary 5.15 we have

THEOREM 5.22. *The tests which reject when* $S_1 \ge c(\alpha)$ *where*

$$P_g\left\{ \int_0^1 Z^2(t)\, dt \ge c(\alpha) \right\} = \alpha$$

asymptotically have level α *for* $H: g = g$ *and are consistent against all* $f \ne g$ *which satisfy* A *and* B.

PROOF. That the tests have level α is immediate from corollary 5.15. We check consistency for S_0.

Note first that if $f \ne g$

$$\int_0^1 (H(t) - t)^2\, dt > 0. \tag{5.23}$$

If not, since $H(e^{-s})$ is the Laplace transform of $\lambda(X_1)$ and equals e^{-s} a.e., then $P_f[\lambda(X_1) = 1] = 1$, implying $f = g$ a.e. Write

$$S_0 = \int_0^1 Z_n^2(t)\, dt + 2\sqrt{n} \int_0^1 Z_n(t)(E_f\hat{H}(t) - E_g\hat{H}(t))\, dt + n \int_0^1 (E_f\hat{H}(t) - E_g\hat{H}(t))^2\, dt.$$

Then

$$\int_0^1 Z_n^2(t)\, dt = O_p(1)$$

$$\sqrt{n} \int_0^1 Z_n(t)(E_f\hat{H}(t) - E_g\hat{H}(t))\, dt = O_p(\sqrt{n})$$

$$n \int_0^1 (E_f\hat{H}(t) - E_g\hat{H}(t))^2\, dt \sim n \int_0^1 (H(t) - t)^2\, dt = O(n)$$

by (5.23). Therefore,

$$S_0 \to_P \infty$$

and consistency follows.

NOTE. In his thesis, M. Schilling (1979) has made a far reaching investigation of the power of this and related tests against contiguous alternatives, has constructed tables of the asymptotic null distribution of S_0 for $m = 1$ and ∞ and has studied the efficiency of the large m and n approximation through simulation.

APPENDIX

In this appendix we give the statements and proofs of several lemmas of a technical or computational nature which are used in the previous sections. We begin with a key lemma due to Stone (1977).

LEMMA S. *For each m and norm $\|\cdot\|$ there exists $\alpha(m) < \infty$ such that it is possible to write R^m as the union of $\alpha(m)$ disjoint cones $C_1, \cdots, C_\alpha$ with 0 as their common peak such that if*

$$x, y \in C_j,\ x, y \neq 0, \qquad \text{then} \quad \|x - y\| < \max(\|x\|, \|y\|), \qquad j = 1, \cdots, \alpha(m).$$

The following straightforward modification of Stone's argument shows that the lemma is valid for any norm.

PROOF. By compactness of the surface of the unit sphere $\partial S(0, 1)$ we can find $\tilde{C}_1, \cdots, \tilde{C}_{\alpha(m)}$ disjoint sets such that,

(i) $\cup_{j=1}^{\alpha(m)} \tilde{C}_j = \partial S(0, 1)$

(ii) $x, y \in \tilde{C}_j \Rightarrow \|x - y\| < 1$.

Let

$$C_j = \{\lambda x : x \in \tilde{C}_j, \lambda \geq 0\}, \qquad j = 1, \cdots, \alpha(m).$$

Suppose $x = \lambda\tilde{x}$, $y = \eta\tilde{y}$, $\tilde{x}, \tilde{y} \in \tilde{C}_j$. Suppose w.l.o.g. $\lambda \leq \eta$. Then,

$$\|x - y\| = \eta\left\|\frac{\lambda}{\eta}\tilde{x} - \tilde{y}\right\| \leq \left\{\left(1 - \frac{\lambda}{\eta}\right)\|\tilde{y}\| + \frac{\lambda}{\eta}\|\tilde{x} - \tilde{y}\|\right\} < \|y\|.$$

The following are easy corollaries of Lemma S.

COROLLARY S1. *For any set of n distinct points, $x_1, \cdots, x_n$ in R^m, x_1 can be the nearest neighbor of at most $\alpha(m)$ points.*

COROLLARY S2. *If $C_1, \cdots, C_{\alpha(m)}$ are as in Lemma S, y_0 is arbitrary, $x \in C_j + y_0$, then*

$$S(x, \|x - y_0\|) \supset S(y_0, \|x - y_0\|) \cap (C_j + y_0).$$

The following consequence of S2 is needed for the proof of Lemma A2 but is of independent interest.

THEOREM A1. *Let Y be a random m vector with distribution G, density g, and let y_0 be a fixed point,*

$$Q = G(S(Y, \|Y - y_0\|)).$$

Then,

$$P[Q \leq q] \leq \alpha(m)q, \qquad 0 \leq q \leq 1. \tag{A.2}$$

PROOF. First let $y_0 = 0$ and let G_j be the conditional distribution of $Y \mid Y \in C_j$ and p_j

$= G(C_j)$, where the C_j are given by corollary S2. Then,

$$P[Q \le q] = \textstyle\sum_j \{p_j P[Q \le q \mid Y \in C_j] : p_j > 0\}. \tag{A.3}$$

But $Y \in C_j$ implies by Corollary S2 that

$$G(S(Y, \| Y \|)) \ge p_j G_j(S(0, \| Y \|) \cap C_j).$$

Hence, for $p_j > 0$.

$$P[Q \le q \mid Y \in C_j] \le P\left[G_j(S(0, \| Y \|)) \le \frac{q}{p_j} \,\middle|\, y \in C_j \right] = \frac{q}{p_j} \tag{A.4}$$

since, given $Y \in C_j$, $G_j(S(0, \| Y \|))$ has a uniform distribution on (0, 1). (A.2) and (A.3) imply (A.1) if $y_0 = 0$. For the general case shift everything by y_0 and apply Corollary S2 in full generality.

COROLLARY A5. *If Q is as in Theorem* A.1, $r \ge 0$

$$E(1 - Q)^r Q \le M(r + 1)^{-2}$$

where M depends only on m.

PROOF. Since $0 \le Q \le 1$ we may w.l.o.g. take $r \ge 2$. By integration by parts

$$E(1 - Q)^r Q = \int_0^1 P[Q \le q]\{-(1 - q)^r + rq(1 - q)^{r-1}\, dq\} \le \alpha(m) r \int_0^1 q^2(1 - q)^{r-1}\, dq$$

$$\le r(r - 1)^{-3}\alpha(m) \int_0^{r-1} w^2\left(1 - \frac{w}{r - 1}\right)^{r-1} dw$$

$$\le 2\alpha(m) r(r - 1)^{-3} \le M(r + 1)^{-2}.$$

We proceed to Lemmas A6 and A10.

LEMMA A6. *Let*

$$F_{i1} = [X_i \ne \tilde{X}_i], \qquad F_{i2} = [X_i = \tilde{X}_i, R_i \ne \tilde{R}_i]; \qquad F_{i3} = [J_i = 2 \text{ or } \tilde{J}_i = 2].$$

Then

$$P[F_{1J}] \le M\left(\frac{r}{n} + F(S)\right), \qquad \forall j. \tag{A.7}$$

$$P[F_{1j} \cap F_{1k}] \le M\left(\frac{r^2}{n^2} + F^2(S)\right), \qquad \forall j \ne k. \tag{A.8}$$

PROOF. All these estimates follow by symmetry arguments as in the proof of Lemma 2.27. We prove one of the estimates of (A.8) as an example. Note that we may without loss of generality take $r \le n/4$ (say). Then

$$\begin{aligned} P[F_{12} \cap F_{13}] &\le [(n - r)(n - r - 1)]^{-1} E[\textstyle\sum_{i=1}^{n-r} I(F_{i2}) \sum_{k=1, k\ne i}^{n-r} (I(J_i = k) + I(\tilde{J}_i = k))] \\ &\le 8\alpha(m) n^{-2} E(N + r) \end{aligned} \tag{A.9}$$

by Corollary S1. But

$$8\alpha(m) n^{-2} E(N + r) \le \frac{M}{n}\left(\frac{r}{n} + F(S)\right) \le M\left(\frac{r^2}{n^2} + F^2(S)\right).$$

Clearly the bounds (A.7) and (A.8) are overestimates in this case. We have written the lemma in this way for compactness.

LEMMA A10. *With the same definitions for $j = 1, 2$,*

$$EI(F_{1j})\frac{N}{n} \le M\left(\frac{r^2}{n^2} + F^2(S)\right) \tag{A.11}$$

$$EI(F_{1j})F(S_1) \le M\left(\frac{r^2}{n^2} + F^2(S)\right) \tag{A.12}$$

$$EI(F_{1j})F(\tilde{S}_1) \le M\left(\frac{r^2}{n^2} + F^2(S)\right). \tag{A.13}$$

PROOF. a) $j = 1$

$$EI(F_{11})\frac{N}{n} = F(S)\left(1 + \left(1 - \frac{r-1}{n}\right)F(S)\right)$$

$$EI(F_{11})F(S_1) = P(F_{11})EF(S_1) = \frac{F(S)}{n}.$$

Let

$$R_i^* = \min\{\|\tilde{X}_i - \tilde{X}_j\| : 1 \le j \le n - r, j \ne i\}.$$

Then,

$$EI(F_{11})F(\tilde{S}_1) \le EI(F_{11})F(S(\tilde{X}_1, R_1^*)) \le (n-r)^{-1}F(S)(1 - F(S)) + F^2(S).$$

The bounds (A.11–A.13) are immediate for $r \le n/4$ and trivial (for large enough M) for $r > n/4$.

b) $j = 2$

$$EI(F_{12})\frac{N}{n} = \left(1 - \frac{r-1}{n}\right)P[F_{12} \cap F_{21}] \le 2\alpha(m)\frac{EN(N+r)}{n(n-r-2)} \le M\left(\frac{r}{n}F(S) + F^2(S)\right)$$

for $r \le n/4$ and (A.11) follows. To prove (A.12) begin by writing,

$$\begin{aligned} EI(F_{12})F(S_1) \le\ & EI(X_1 = \tilde{X}_1, R_1 \le \tilde{R}_1)F(S_1) + EI(X_1 = \tilde{X}_1, R_{10} > R_{1c})F(S_1) \\ & + \textstyle\sum_{j=1}^{r} EI(X_1 = \tilde{X}_1, R_{10} > \|X_1 - x_j\|)F(S_1) \end{aligned} \tag{A.14}$$

where,

$$R_{10} = \min\{\|\tilde{X}_j - \tilde{X}_1\| : X_j = \tilde{X}_j, j \ne 1, 1 \le j \le n - r\}$$

$$R_{1c} = \min\{\|\tilde{X}_j - \tilde{X}_1\| : X_j \ne \tilde{X}_j, j \ne 1, 1 \le j \le n - r\}.$$

Then, we bound

$$EI(X_1 = \tilde{X}_1, R_1 < \tilde{R}_1)F(S_1) < EI(X_{J_1} \ne \tilde{X}_{J_1})F(S_1) = n^{-1}F(S). \tag{A.15}$$

Next,

$$\begin{aligned} & EI(X_1 = \tilde{X}_1, R_{10} > R_{1c})F(S_1) \\ & \le E\{P[F_S(S(\tilde{X}_1, R_{10})) > F_s(S(\tilde{X}_1, R_{1c})) \mid N, \tilde{X}_1, R_{1c}, X_1 = \tilde{X}_1] \\ & \quad \cdot F_S(S(\tilde{X}_1, R_{1c}))I(X_1 = \tilde{X}_1)\} \\ & = E[(1 - F_S(S(\tilde{X}_1, R_{1c})))^{K-1}F_S(S(\tilde{X}_1, R_{1c}))I(X_1 = \tilde{X}_1)] \end{aligned} \tag{A.16}$$

where $K = n - r - N$

$$\le EN\int_0^1 (1-w)^{n-r-2}w\,dw = [(n-r)(n-r-1)]^{-1}EN$$

$$\le M\frac{r}{n}F(S)$$

for $r \le n/4$.

The next to last inequality follows since, given $X_1 = \tilde{X}_1$ and N, $F_S(\tilde{X}_1, R_{1c})$ is distributed as the minimum of N uniform (0, 1) variables. Finally, arguing as above,

$$\text{(A.17)}\quad \begin{aligned} &EI(X_1 = \tilde{X}_1, R_{10} > \| X_1 - x_j \|)F(S_1) \\ &\quad \le E(1 - F_S(S(\tilde{X}_1, \| \tilde{X}_1 - x_j \|)))^{K-1} F_S(S(\tilde{X}_1, \| \tilde{X}_1 - x_j \|))I(X_1 = \tilde{X}_1). \end{aligned}$$

Given $X_1 = \tilde{X}_1$, we can apply Corollary A.1 noting that $F_S(S(\tilde{X}_1, \| \tilde{X}_1 - x_j \|))$ has the distribution of Q with $G = F_S$, $x_j = y_0$. Since conditionally $K - 1$ has a binomial $(n - r - 1, 1 - F(S))$ distribution, we obtain as a bound for (A.17),

$$\text{(A.18)}\qquad ME(K^{-2} \mid X_1 = \tilde{X}_1) \le 3/2 M(1 - F(S))^{-2}(n - r)^{-2}.$$

Therefore, we obtain

$$\text{(A.19)}\qquad \textstyle\sum_{j=1}^{r} EI(X_1 = \tilde{X}_1, R_{10} > \| X_1 - x_j \|)F(S_1) \le M\left(\dfrac{r^2}{n^2} + F(S)\right)$$

for

$$r \le \frac{n}{4}, \qquad F(S) \le \frac{1}{4} \ \text{(say)}.$$

Combining (A.15), (A.16) and (A.17) we obtain (A.12) for $j = 2$, since the restrictions on r and F can be absorbed into M for the final bound. Finally,

$$\text{(A.20)}\quad EI(F_{12})\tilde{S}_1 \le EI(X_1 = \tilde{X}_1, \tilde{R}_1 \le R_1)F(S_1) + EI(X_1 = \tilde{X}_1, X_{J_1} \ne \tilde{X}_{J_1})F(S(\tilde{X}_1, R_{10})).$$

The first term in (A.20) has been bounded in (A.14) and (A.19). The second is bounded as in (A.15) by

$$F(S)E\left(\frac{1}{K} \,\middle|\, X_1 = \tilde{X}_1\right) \le MF(S)\frac{r}{n}, \qquad F(S) \le \frac{1}{4},$$

$r \le n/4$. (A.13) follows for $j = 2$ and the lemma is proved.

REFERENCES

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

LOFTSGAARDEN, D. O. and QUESENBERRY, C. P. (1965). A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.* **36** 1049-1051.

MORRIS, C. (1975). Central limit theorems for multinomial sums. *Ann. Statist.* **3** 165–188.

ROGERS, W. (1977). Thesis, Stanford University.

SCHILLING, M. (1979). Thesis, University of California, Berkeley.

SHORACK, G. (1973). Convergence of reduced empirical and quantile processes, etc. *Ann. Statist.* **1** 146–152.

STONE, C. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645.

STATISTICS DEPARTMENT
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 12128–12131, December 1996
Statistics

What is a linear process?

(chaos plus noise/ergodicity/Gaussian process/infinitely divisible law/nonlinear process)

PETER J. BICKEL AND PETER BÜHLMANN

Department of Statistics, University of California, Berkeley, CA 94720-3860

Contributed by Peter J. Bickel, August 1, 1996

ABSTRACT **We argue that given even an infinitely long data sequence, it is impossible (with any test statistic) to distinguish perfectly between linear and nonlinear processes (including slightly noisy chaotic processes). Our approach is to consider the set of moving-average (linear) processes and study its closure under a suitable metric. We give the precise characterization of this closure, which is unexpectedly large, containing nonergodic processes, which are Poisson sums of independent and identically distributed copies of a stationary process. Proofs of these results will appear elsewhere.**

1. Preliminary Description of Problems and Results

It has long been known, though perhaps not always appreciated, that it is impossible to test whether a set of observations comes from a "linear" ergodic or nonergodic Gaussian process since any nonergodic Gaussian process can be arbitrarily well approximated in a suitable metric by ergodic Gaussian processes, which are necessarily linear. We will present here a novel result that essentially any stationary process cannot be sharply distinguished from a linear process. Loosely, we consider the following problem: Given a partial realization $x_1, \ldots, x_n$ of a strictly stationary stochastic process $\{X_t\}_{t\in\mathbb{Z}}$, where $\mathbb{Z} = \{0, \pm1, \pm2, \ldots\}$, when can we conclude that the process is linear?

In recent years there has been a considerable interest in nonlinear time series analysis in the statistical, econometric, and engineering literatures (1–3). "Nonlinear" corresponds to many subclassifications, such as "bilinear" or "threshold autoregressive." Also, noisy chaotic processes defined by

$$X_t = f(X_{t-1}) + \varepsilon_t \ (t \in \mathbb{Z}),$$

where ε_t i.i.d. with $\mathbf{E}[\varepsilon_t] = 0$ and $f: \mathbb{R} \to \mathbb{R}$, define a subclass of nonlinear processes (for general f). But, at least linearity is fairly unambiguously specified. A linear stationary process $(X_t)_{t\in\mathbb{Z}}$ is usually described by

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} \ (t \in \mathbb{Z}), \qquad \textbf{[1.1]}$$

where ε_t i.i.d. with $\mathbf{E}[\varepsilon_t] = 0$, $\mathbf{E}|\varepsilon_t|^2 < \infty$ and $\sum_{j=0}^{\infty}\psi_j^2 < \infty$. Such processes are also called moving-average (MA) processes. Here, we always assume existence of second moments. There is no loss of generality in assuming $\mathbf{E}[X_t] = 0$. Note that causal (minimum phase) autoregressive or ARMA processes are also representable as MA processes.

Given a finite stretch of a realization of a stationary process, one can try and test the hypothesis of linearity as stated in Eq. **1.1**. Such omnibus tests have been proposed, mainly by looking at higher order spectra (4, 5). But this hypothesis can be rejected only if alternatives are not well approximated by processes satisfying the hypothesis. The problem of testing H_0 about MA representation as in Eq. **1.1** leads then to the problem of studying the closure of the set of probability distributions of MA processes as given in Eq. **1.1** (MA closure).

The notion of a closed set requires the specification of a topology. We work here with the Mallows metric (6), also known as the Wasserstein metric, and with the stronger total variation metric. (For details, see sections 2.1 and 2.2.) We always identify real-valued stochastic processes, indexed by $\mathbb{Z}$, with their corresponding probability distributions; we then prefer to state our results in terms of stochastic processes.

We will argue in section 1.3 that the Mallows MA closure is exhausted by three types of processes. The first type is the set of stationary Gaussian processes with mean zero, i.e.,

$$S_1 = \{(X_t)_{t\in\mathbb{Z}};\ (X_t)_{t\in\mathbb{Z}} \text{ stationary Gaussian process with } \mathbf{E}[X_t] = 0\}.$$

The second type is the set of genuine MA processes, i.e.,

$$S_2 = \{(X_t)_{t\in\mathbb{Z}};\ X_t \text{ as defined in Eq. } \mathbf{1.1}\}.$$

The third type which arises is more surprising. We essentially can get Poisson sums of independent and identically distributed copies of stationary processes in the following sense. Denote by

$$(\xi_{t;1})_{t\in\mathbb{Z}},\ (\xi_{t;2})_{t\in\mathbb{Z}},\ \ldots,$$

a sequence of independent, real-valued, stationary processes with mean zero and finite second moments $\mathbf{E}|\xi_{t;1}|^2 = \sigma^2_{\xi;1}$, $\mathbf{E}|\xi_{t;2}|^2 = \sigma^2_{\xi;2}, \ldots$. Moreover, we construct for every $i \in \mathbb{N} = \{1, 2, \ldots\}$ a sequence of independent copies of $(\xi_{t;i})_{t\in\mathbb{Z}}$, namely

$$(\xi_{t;i,1})_{t\in\mathbb{Z}},\ (\xi_{t;i,2})_{t\in\mathbb{Z}},\ \ldots.$$

Thus we have constructed a sequence of processes

$$\{(\xi_{t;i,j})_{t\in\mathbb{Z}}\}_{i,j\in\mathbb{N}} \text{ independent processes over the index set}$$

$$i, j \in \mathbb{N}, (\xi_{t;i,1})_{t\in\mathbb{Z}}, (\xi_{t;i,2})_{t\in\mathbb{Z}}, \ldots \text{ i. i. d.}, \mathbf{E}[\xi_{t;i,j}] = 0, \mathbf{E}|\xi_{t;i,j}|^2 = \sigma^2_{\xi;i} < \infty. \qquad \textbf{[1.2]}$$

Let

$$N_1, N_2, \ldots \text{ independent, } N_i \sim \text{Poisson}(\lambda_i),\ \lambda_i \geq 0 \text{ for all } i \in \mathbb{N}. \qquad \textbf{[1.3]}$$

Then the third type is given by the following set of processes,

$$S_3 = \left\{ (X_t)_{t\in\mathbb{Z}};\ X_t = \sum_{i=1}^{\infty} \sum_{j=1}^{N_i} \xi_{t;i,j},\ (\xi_{t;i,j})_{t\in\mathbb{Z}},\ N_i \text{ satisfying } \mathbf{1.2}, \mathbf{1.3}, \text{ and} \sum_{i=1}^{\infty} \lambda_i \sigma^2_{\xi;i} < \infty \right\}.$$

Abbreviation: MA, moving average.

We make the convention that $\sum_{j=1}^{0} \xi_{t;i,j} = 0$. Elements of S_3 are typically nonergodic processes whose finite dimensional distributions are infinitely divisible non-Gaussian.

1.1. Nonergodic Limits and Separation Dilemma. In an informal way, the ergodic hypothesis postulates the equality of time-averages with averages over the elements in a probability space (in statistical mechanics, "phase-averages" in the phase space of a mechanical system). But the distinction between ergodic and nonergodic processes can be blurred.

Example 1.1: Consider the sequence of finite order MA processes,

$$X_t^{(n)} = \sum_{j=1}^{n} \xi_{j;1} U_{t-j,n} Z_{t-j,n} \ (t \in \mathbb{Z}),$$

with U_t i.i.d., $\mathbb{P}[U_t = 1] = 1 - \mathbb{P}[U_t = 0] = \lambda/n$ $(\lambda > 0)$, Z_t i.i.d. $\sim t_5$, Student's t distribution with 5 degrees of freedom, and coefficients $(\xi_{j;1})_{j\in\mathbb{N}}$ which are a fixed realization of the Gaussian AR(1), $\xi_{j;1} = 0.9\xi_{j-1;1} + \eta_j$, η_j i.i.d. $\sim \mathcal{N}(0, 1)$.

For every $n \in \mathbb{N}$, these are ergodic MA processes of finite order n. But they exhibit a behavior which can be interpreted as nonergodic and "nonstationary," and which seems far from what one expects of a linear process. The reason is that they are close to a nonergodic member in S_3.

To illustrate the nonergodic phenomenon, we show in Fig. 1*A* nine realizations of sample size 500 of the process in Example 1.1 with $n = 50$. Fig. 1*A* tells in a quite impressive manner how different such realizations can be, and thus indicates that time-averages are not compatible with phase-averages over different realizations—i.e., nonergodic behavior. Fig. 1*B* shows one realization of sample size 5000 of the MA process in Example 1.1 with $n = 200$, now indicating nonstationarity. Different stretches of the sequence exhibit very different behaviors. This is the typical pattern for a time series with innovation outliers (7). Indeed, our model is an extreme case with innovations being either zero with probability $1 - \lambda/n$ or being a realization from a long-tailed distribution with probability λ/n. Note that outliers are with reference to the Gaussian distribution; it is the nonnormality of innovations which can lead to MA processes being close to a process in S_3.

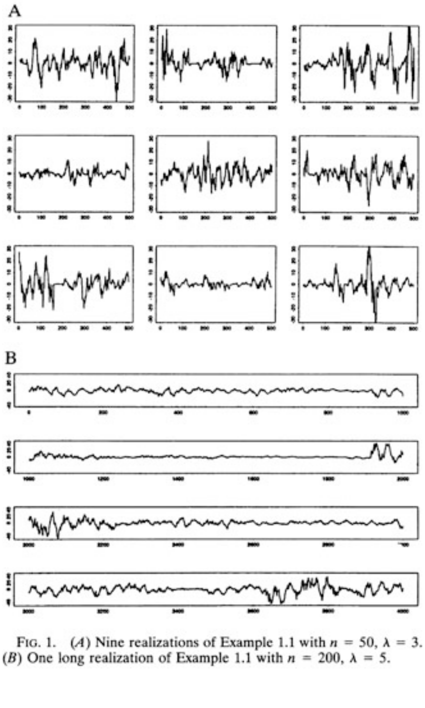


FIG. 1. (*A*) Nine realizations of Example 1.1 with $n = 50$, $\lambda = 3$. (*B*) One long realization of Example 1.1 with $n = 200$, $\lambda = 5$.

Example 1.1 is a special case of a very disturbing subclass of MA processes close to S_3. Given any, even infinitely long, realization $(\xi_t)_{t\in\mathbb{Z}}$ from any stationary process, consider the process $(X_t)_{t\in\mathbb{Z}} \in S_3$, where $X_t = \sum_{j=1}^{N} \xi_{t;j}$ $(t \in \mathbb{Z})$ with $N \sim$ Poisson(1) and $(\xi_{t;1})_{t\in\mathbb{Z}} = (\xi_t)_{t\in\mathbb{Z}}$, $(\xi_{t;2})_{t\in\mathbb{Z}}, \ldots$ independent identically distributed copies. It can be shown that this process is an element of the MA closure, compare also with Fact 1.4 in section 1.3. Since $\mathbb{P}[N = 1] = e^{-1} > 0.36$, we obtain $\mathbb{P}[X_t = \xi_t \text{ for all } t \in \mathbb{Z} | (\xi_t)_{t\in\mathbb{Z}}] > 0.36$. Summarizing, we have the following separation dilemma.

FACT 1.1. *Given any stationary process $(\xi_t)_{t\in\mathbb{Z}}$, there exists a nonergodic, stationary process $(X_t)_{t\in\mathbb{Z}}$ in the MA closure, which is an element of S_3 and has with positive probability exactly the same sample path as $(\xi_t)_{t\in\mathbb{Z}}$. More precisely,*

$$\mathbb{P}[X_t = \xi_t \textit{ for all } t \in \mathbb{Z} | (\xi_t)_{t\in\mathbb{Z}}] > 0.36 \textit{ almost surely.}$$

Details are given in Theorem 2.2. This separation dilemma is of the same nature as de Finetti's Theorem which can be thought of as stating the impossibility of distinguishing exchangeable from i.i.d. sequences (8, pp. 40–42).

In terms of the whole stochastic process, rather than a sample path, we have the following.

FACT 1.2. *The MA closure does not contain the set of ergodic, stationary processes.*

To show the validity of Fact 1.2, it is sufficient to give an example.

Example 1.2: Consider the stationary Markov chain $(X_t)_{t\in\mathbb{Z}}$, given by $X_t \in \{0, 1\}$ with $\mathbb{P}[X_1 = 0] = \mathbb{P}[X_1 = 1] = 1/2$, $\mathbb{P}[X_1 = 0|X_0 = 0] = \mathbb{P}[X_1 = 0|X_0 = 1] = \pi$, $0 < \pi < 1/2$. Then $(X_t)_{t\in\mathbb{Z}}$ is ergodic. Moreover, the probability distribution of X_t is not divisible, since the convolution of two nondegenerate distributions would place mass on at least three points, whereas X_t is only binary. Hence, the distribution of X_t cannot be approximated by any MA process and $(X_t)_{t\in\mathbb{Z}}$ can therefore not be an element of the MA closure.

It is possible to construct an ergodic, stationary process, with marginal distributions having a density with respect to Lebesgue measure, which is not an element of the MA closure (see ref. 15).

There are probably many ergodic, stationary processes, which are not elements of the MA closure. A possible candidate is the bilinear process, given by

$$X_t = -0.4X_{t-1} + 0.4X_{t-1}\varepsilon_{t-1} + \varepsilon_t (t \in \mathbb{Z}),$$

where ε_t i.i.d. $\sim \mathcal{N}(0, 1)$ (see figure 3.10 in ref. 9).

This process is stationary and ergodic (10). It is also immediate that the process is non-Gaussian. As argued in Subba Rao and Gabr (table 3.2 and figure 3.3 in ref. 9), this bilinear process is not representable as a moving average process. However, the MA closure also contains the class S_3 some of whose members may be ergodic.

1.2. The Testing Dilemma. There is considerable interest in testing the hypothesis that an observed time series is a linear process. Several authors propose different procedures for testing the hypothesis of MA representation (4, 5) and of autoregressive representation (11).

Consider the problem of distinguishing between the hypothesis $H_0 : (X_t)_{t\in\mathbb{Z}}$ is a linear process against the alternative H_A : $(X_t)_{t\in\mathbb{Z}}$ is a specific stationary process (not approximable by H_0 processes). Do there exist critical regions C_n for rejecting H_0, such that $\mathbb{P}_{H_0}[(X_1, \ldots, X_n) \in C_n] \to \alpha > 0$ and $\mathbb{P}_{H_A}[(X_1, \ldots, X_n) \notin C_n] \to 0$ as $n \to \infty$. That is, can one distinguish perfectly between H_0 and H_A at any level of significance α? Fact 1.1 can be restated as follows.

FACT 1.3. *In testing the hypothesis* H_0 *about MA representation against any fixed one-point alternative* H_A *about a nonlinear, stationary process, there is no test with asymptotic significance level* $\alpha < 0.36$ *having limiting power 1 as the sample size tends to infinity.*

1.3. Exhausting the MA Closure. The sets S_1, S_2, S_3 are not rich enough to exhaust the Mallows MA closure. To achieve this, we need sums of processes of the different types. We introduce an adding operation for processes and define

$(X_t)_{t\in\mathbb{Z}} \oplus (Y_t)_{t\in\mathbb{Z}}$ is the process $(X_t + Y_t)_{t\in\mathbb{Z}}$,

where the processes $(X_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$ are independent.

We then set

$$S_i \oplus S_j = \{(X_t)_{t\in\mathbb{Z}} \oplus (Y_t)_{t\in\mathbb{Z}};\ (X_t)_{t\in\mathbb{Z}} \in S_i,\ (Y_t)_{t\in\mathbb{Z}} \in S_j\},\ i, j \in \{1, 2, 3\},$$

and make the common convention that all S_i $(i = 1, 2, 3)$ also contain the null element $X_t \equiv 0$ for all $t \in \mathbb{Z}$.

The representation of a process as a $\oplus$-sum of elements in S_i $(i = 1, 2, 3)$ is not unique even in the Gaussian case.

FACT 1.4. *The closure of the set of MA processes is given by*

$$\{S_1 \oplus S_2\} \cup \{S_1 \oplus S_3\}.$$

Details are given in Theorem 2.1. Mallows (12) argues that a linear process such as in Eq. **1.1** is close to a Gaussian process if $\max_{j\geq 0} |\psi_j|$ is small. This is no longer true if one considers sequences $\{(X_{t,n})_{t\in\mathbb{Z}}\}_{n\in\mathbb{N}}$ of linear processes with coefficients $\psi_{j,n}$ as above and variables $\varepsilon_{t,n}$, which are i.i.d. but depend on n. Then, if $\max_{j\geq 0}|\psi_{j,n}| \to 0$ $(n \to \infty)$ the process $(X_{t,n})_{t\in\mathbb{Z}}$ can have marginal distributions close to a non-Gaussian (not purely Gaussian) infinitely divisible law. Our result is in the spirit of Lévy (13) and uses his arguments. He showed that every continuous time process $(X_t)_{t\in\mathbb{Z}}$ with independent increments must have an infinitely divisible law and that such processes can be realized by a process with independent time homogeneous increments.

2. Precise Formulations

We consider real-valued, stationary processes $(X_t)_{t\in\mathbb{Z}}$ with expectation zero and finite variances. Thus, an appropriate probability space is $(\mathbb{R}^{\mathbb{Z}}, \mathcal{B}, \mathcal{P})$, where $\mathcal{B}$ denotes the Borel σ-field on $\mathbb{R}^{\mathbb{Z}}$ and $\mathcal{P}$ a class of stationary probability measures on $(\mathbb{R}^{\mathbb{Z}}, \mathcal{B})$, such that for every $P \in \mathcal{P}$,

$$\mathbf{E}_P[X] = \int_{\mathbb{R}} x\,d(P \circ \pi_0^{-1})(x) = 0,$$

$$\mathbf{E}_P|X|^2 = \int_{\mathbb{R}} x^2 d(P \circ \pi_0^{-1})(x) < \infty,$$

where $\pi_{t_1}, \ldots, {}_{t_m} : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^m$, $(x_t)_{t\in\mathbb{Z}} \mapsto (x_{t_1}, \ldots, x_{t_m})$, $t_1, \ldots, t_m \in \mathbb{Z}$.

We always identify a probability measure $P \in \mathcal{P}$ with its corresponding real-valued stochastic process.

It is possible to metrize the space $\mathcal{P}$ with a metric d (see sections 2.1 and 2.2). The closure with respect to the metric d of sets in $\mathcal{P}$, or equivalently of stationary real-valued stochastic processes with distributions in $\mathcal{P}$, is defined in the usual topological sense. We are particularly interested in the closure of MA processes (MA closure). Thus, we will consider sequences

$$\left\{\left(X_{t,n} = \sum_{j=0}^{\infty} \psi_{j,n}\varepsilon_{t-j,n}\right)_{t\in\mathbb{Z}}\right\}_{n\in\mathbb{N}}. \qquad [2.1]$$

2.1. Mallows Metric. We define the Mallows metric d_2 on $\mathcal{P}$, by

$$d_2(P_1, P_2) = \sum_{m=1}^{\infty} d_2^{(m)}(P_1 \circ \pi_{1,\ldots,m}^{-1}, P_2 \circ \pi_{1,\ldots,m}^{-1})2^{-m}, \quad P_1, P_2 \in \mathcal{P},$$

where $d_2^{(m)}(P_1 \circ \pi_{1,\ldots,m}^{-1}, P_2 \circ \pi_{1,\ldots,m}^{-1}) = \inf\{(\mathbf{E}\|X - Y\|^2)^{1/2}\}$ when the infimum is taken over all jointly distributed $(X, Y) \in \mathbb{R}^{2m}$ having marginals $P_1 \circ \pi_{1,\ldots,m}^{-1}$ and $P_2 \circ \pi_{1,\ldots,m}^{-1}$; $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^m$.

The following characterization is useful. Let $P_n, P \in \mathcal{P}$ and denote by $\Rightarrow$ weak convergence of probability measures. Then,

$$d_2(P_n, P) \to 0\ (n \to \infty)$$

is equivalent to the following two statements

$$P_n \circ \pi_{t_1,\ldots,t_m}^{-1} \Rightarrow P \circ \pi_{t_1,\ldots,t_m}^{-1}\ (n \to \infty) \text{ for all } t_1, \ldots, t_m \in \mathbb{Z}, m \in \mathbb{N},$$

$$\int_{\mathbb{R}} x^2 d(P_n \circ \pi_0^{-1})(x) \to \int_{\mathbb{R}} x^2 d(P \circ \pi_0^{-1})(x)\ (n \to \infty),$$

that is, all finite dimensional distributions at $t_1, \ldots, t_m$ converge weakly and the variance of the marginal at any time point t converges (see ref. 14). We also use the notation for the corresponding processes, $d_2((X_{t,n})_{t\in\mathbb{Z}}, (X_t)_{t\in\mathbb{Z}}) = d_2(P_n, P)$, where $(X_{t;n})_{t\in\mathbb{Z}} \sim P_n$, $(X_t)_{t\in\mathbb{Z}} \sim P$.

2.2. Variation Metric. The question about distinguishing perfectly between two stationary processes requires a stronger metric than the Mallows d_2. The variation metric allows a precise formulation.

As before, let $P_1, P_2 \in \mathcal{P}$ and define the variation metric as

$$d_V(P_1, P_2) = \sum_{m=1}^{\infty} d_V^{(m)}(P_1 \circ \pi_{1,\ldots,m}^{-1}, P_2 \circ \pi_{1,\ldots,m}^{-1})2^{-m},$$

where $d_V^{(m)}(P_1 \circ \pi_{1,\ldots,m}^{-1}, P_2 \circ \pi_{1,\ldots,m}^{-1}) = \sup\{|P_1 \circ \pi_{1,\ldots,m}^{-1}[A] - P_2 \circ \pi_{1,\ldots,m}^{-1}[A]|; A \in \mathcal{B}(\mathbb{R}^m)\}$, $\mathcal{B}(\mathbb{R}^m)$ the Borel σ-algebra of $\mathbb{R}^m$. This definition reflects the nonuniform convergence of finite dimensional distributions in the variation metric. Here we do not require convergence of second moments. Distinguishing perfectly is characterized as follows. Let P_1, P_2 be ergodic probability measures in $\mathcal{P}$. Then

$d_V(P_1, P_2) > 0$ if and only if there exist test functions

$$\varphi_m : \mathbb{R}^m \to \mathbb{R}, 0 \leq \varphi_m \leq 1, \text{ such that } \mathbf{E}_{P_1}[\varphi_m(X_1, \ldots, X_m)] \to 0, \mathbf{E}_{P_2}[\varphi_m(X_1, \ldots, X_m)] \to 1 (m \to \infty).$$

2.3. Closure for MA Processes. We consider first the Mallows d_2 closure for MA processes, that is, sequences as defined in Eq. **2.1**. Without loss of generality we can scale the innovations and assume: (A): For every $n \in \mathbb{N}$, $(\varepsilon_{t,n})_{t\in\mathbb{Z}}$ is an i.i.d. sequence with

$$\mathbf{E}[\varepsilon_{t,n}] = 0,\ \mathbf{E}|\varepsilon_{t,n}|^2 = 1.$$

The following result describes the Mallows MA closure.

THEOREM 2.1. (i) *Consider a sequence of MA processes as defined in Eq.* **2.1** *converging in the* d_2 *sense, satisfying* (A) *and one of the following:*

(A1): $d_2^{(1)}(\varepsilon_{t,n}, \varepsilon_t) \to 0$ $(n \to \infty)$, *where* $(\varepsilon_t)_{t\in\mathbb{Z}}$ *is an i.i.d. sequence with* $\mathbf{E}[\varepsilon_t] = 0$.

(A2): $max_{j\geq 0}|\psi_{j,n}| \to 0$ $(n \to \infty)$.

Then, the d_2 *limit of such a sequence is in* $\{S_1 \oplus S_2\} \cup \{S_1 \oplus S_3\}$.

(ii) *Every element of* $\{S_1 \oplus S_2\} \cup \{S_1 \oplus S_3\}$ *can be obtained as a* d_2 *limit of a sequence of MA processes as defined in Eq.* **2.1,** *satisfying* (A) *and* (A1) *or* (A2).

Example 1.1 describes a sequence of MA processes with d_2 limit in S_3. This example can be modified so that the sequence of MA processes also converges in the variation metric to a d_V limit in S_3. This is needed in the following theorem, which has as a consequence that we can never distinguish perfectly between any stationary processes and MA processes even though there are such processes that cannot be approximated arbitrarily closely by MA processes.

THEOREM 2.2. *The MA closure with respect to the variation metric* d_V *has the following features.*

(i) *Let* $(\xi_t)_{t\in\mathbb{Z}}$ *be any stationary process such that for all* $m \in \mathbb{N}$, *the distributions of* $(\xi_1, \ldots, \xi_m)$ *have densities with respect to Lebesgue measure. Then, there exists a process* $(X_t)_{t\in\mathbb{Z}} \in S_3$, *which is an element of the MA closure with respect to the variation metric* d_V, *such that*

$$\mathbb{P}[X_t = \xi_t \textit{ for all } t \in \mathbb{Z}|(\xi_t)_{t\in\mathbb{Z}}] > 0.36 \textit{ almost surely.}$$

(ii) *There exist ergodic, stationary processes as in* (i) *which are not elements of the MA closure with respect to the variation metric* d_V.

The proofs of Theorem 2.1 and 2.2 are given in Bickel and Bühlmann (15). We have looked here at MA processes of infinite order. All our results are also true for sequences of finite (generally unbounded) order MA processes, which are more common in statistical modeling.

3. Discussion

The basic implication of our results is that any stationary process cannot be sharply distinguished from a high enough order MA process. Our proofs in Bickel and Bühlmann (15) show that a high order is a necessity to approximate an arbitrary ergodic, stationary process in the sense of Fact 1.1 and Theorem 2.2.

However, as can be noted from Fig. 1 the phenomenon is quite noticeable even for ratios of number of parameters to observations as low as 0.1. Note that purely chaotic processes do not fall under Theorem 2.2 since $(\xi_1, \ldots, \xi_m)$ do not have a density for m sufficiently large. However, by adding an arbitrarily small amount of white noise to any stationary process including purely chaotic ones we produce a process which can not be distinguished perfectly from an MA process of high order.

We thank David Freedman for helpful comments. P.J.B. was supported in part by National Security Agency Grant MDA 904-94-H-2020 and National Science Foundation Grant 95049555. P.B. was supported in part by the Swiss National Science Foundation.

1. Tong, H. (1990) *Non-Linear Time Series: A Dynamical System Approach* (Oxford Univ. Press, New York).
2. Priestley, M. B. (1988) *Non-Linear and Non-Stationary Time Series Analysis* (Academic, London).
3. Gershenfeld, N. A. & Weigend, A. S. (1994) *Time Ser. Prediction: Forecasting the Future and Understanding the Past* (Addison–Wesley, Reading, PA).
4. Subba Rao, T. & Gabr, M. M. (1980) *J. Time Ser. Anal.* **1,** 145–158.
5. Hinich, M. (1982) *J. Time Ser. Anal.* **3,** 169–176.
6. Mallows, C. L. (1972) *Ann. Math. Stat.* **43,** 508–515.
7. Kleiner, B., Martin, R. D. & Thomson, D. J. (1979) *J. R. Stat. Soc. B* **41,** 313–351.
8. Hartigan, J. A. (1983) *Bayes Theory* (Springer, New York).
9. Subba Rao, T. & Gabr, M. M. (1984) *An Introduction to Bispectral Analysis and Bilinear Time Series Models*, Lecture Notes in Statistics (Springer, Heidelberg), Vol. 24.
10. Akamanam, S. I., Rao, M. Bhaskara & Subramanyam, K. (1986) *J. Time Ser. Anal.* **7,** 157–163.
11. Hjellvik, V. & Tjøstheim, D. (1995) *Biometrika* **82,** 351–368.
12. Mallows, C. L. (1967) *J. Appl. Prob.* **4,** 313–329.
13. Lévy, P. (1948) *Processus Stochastiques et Mouvement Brownien* (Gauthier-Villars, Paris).
14. Bickel, P. J. & Freedman, D. A. (1981) *Ann. Stat.* **9,** 1196–1217.
15. Bickel, P. J. & Bühlmann, P. (1997) *J. Theor. Probab.*, in press.