Klaus Nordhausen · Sara Taskinen
Editors

Modern Nonparametric, Robust and Multivariate Methods

Festschrift in Honour of Hannu Oja

Springer

Modern Nonparametric, Robust and Multivariate Methods

Klaus Nordhausen • Sara Taskinen
Editors

Modern Nonparametric, Robust and Multivariate Methods

Festschrift in Honour of Hannu Oja

Springer

Editors
Klaus Nordhausen
Department of Mathematics and Statistics
University of Turku
Turku, Finland

School of Health Sciences
University of Tampere
Tampere, Finland

Sara Taskinen
Department of Mathematics and Statistics
University of Jyväskylä
Jyväskylä, Finland

ISBN 978-3-319-22403-9 ISBN 978-3-319-22404-6 (eBook)
DOI 10.1007/978-3-319-22404-6

Library of Congress Control Number: 2015951385

Mathematics Subject Classification (2010): 62H10, 62H12, 62H15, 62H20, 62G10, 62G35, 92C55

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Foreword

Hannu Oja has had an extensive and illustrious career. Many things are striking when you look back over the roughly 35 years of his scholarship. His work forms a unity and coherence, focusing on multivariate invariant and equivariant statistical methods. His work steadily evolved from his early results on defining an affine equivariant median, now referred to as the Oja median, in the early 1980s to an efficiently computable affine equivariant median using transform–retransform methods developed in the late 1990s.

It has been my experience, and I am sure I speak for his many students and coauthors that he is the ideal collaborator. He always has insightful comments, is an excellent listener, and is generous to a fault. It has been a privilege and pleasure to work with him. Over a period of many years, I had the further pleasure of visiting Oulu, Tampere, and Jyväskylä to work on problems with Hannu as well as working with and knowing two of his finest students, Jyrki Möttönen and Esa Ollila, both of whom have gone on to impressive research careers.

My interactions with Hannu began in the summer of 1989 when he visited Penn State for a week with Jukka Nyblom as they were on their way to an IMS meeting in Colorado. During that week, we combined some results of their work and some joint work of mine with Bruce Brown into a paper, On Certain Bivariate Tests and Medians, published in 1992 in the *Journal of the American Statistical Association*. Thus began a fruitful long-term relationship. In January 1991, Hannu brought his family to State College for 6 months and was a visiting professor in the statistics department at Penn State. It was not all work during that time. On spring break in March we all went on a driving trip, exploring the eastern part of the USA through Kentucky and then to Florida where in fine Finnish fashion the family could not wait to swim in the cold Atlantic Ocean.

My overall goal here is to discuss the evolution of his research program for the development of multivariate methods beginning in the early 1980s with the publication of his 1983 paper, Descriptive Statistics for Multivariate Distributions, published in *Statistics and Probability Letters* and finish with some brief remarks on his current research interest in invariant coordinate selection (ICS) and independent component analysis (ICA). Along the way, I will mention a few of the initial papers

that he published when moving into new research areas. Elsewhere in this volume there is a thorough analysis of his coauthors and a list of his publications. His vita will reveal an even wider scholarly effort, including consulting in biomedical research and signal processing.

The 1983 descriptive statistics paper is based on the idea of defining a median by minimizing a sum of simplices determined by data points along with a parameter. In a series of papers, this basic idea was expanded to include affine invariant and equivariant sign and rank tests and estimates for various experimental designs. Computation, especially for high-dimensional data, remained problematic. This work is nicely summed up in his 1999 review paper, Affine Invariant Multivariate Sign and Rank Tests and Corresponding Estimates: A Review, published in the *Scandinavian Journal of Statistics*.

A breakthrough occurred when he combined work on non-affine spatial methods developed in a 1995 paper with Jyrki Möttönen entitled Multivariate Spatial Sign and Rank Methods published in the *Journal of Nonparametric Statistics* with transform-retransform methods discussed in a 1998 paper written with Biman Chakraborty and Probal Chaudhuri entitled Operating Transformation-Retransformation on Spatial Median and Angle Test in *Statistica Sinica*. The result was a computationally efficient set of affine invariant and equivariant statistical methods. This work was further refined and elaborated in a 2004 paper with Ron Randles in *Statistical Science* entitled Multivariate Nonparametric Tests. These ideas are at the heart of his seminal 2010 monograph, Multivariate Nonparametric Methods with R.

There is much of interest in this monograph. For example, there is an extensive discussion and development of scatter matrices, another of his research threads. Scatter matrices form the foundation of much of his current interests in ICA and ICS. An early work is his 2006 paper with Seija Sirkiä and Jan Eriksson entitled Scatter Matrices and Independent Component Analysis in the *Austrian Journal of Statistics*. In 2009, the paper Invariant Co-ordinate Selection published in the *Journal of the Royal Statistical Society, Series B* and written with David Tyler, Frank Critchley, and Lutz Dümbgen greatly expanded the framework for this area and brought it to the attention of many more researchers in statistics. Currently, he has ongoing research projects involving the extension and application of ICA to time series and functional data.

From 2008 to 2012, he was an Academy Professor in Finland, a richly deserved honor. If you consider the number of coauthors and their various countries and affiliations, you would conclude that Hannu is a major ambassador for Finland.

State College, PA, USA Tom Hettmansperger
May 2015

Preface

This Festschrift contains a collection of articles dedicated to Hannu Oja, Professor of Statistics at the University of Turku, on the occasion of his 65th birthday.

Hannu can be regarded as one of the most influential statisticians in Finland. His research career has been exceptional. Hannu has served as Professor in the Universities of Jyväskylä, Tampere and Turku, and as Visiting Professor at the Pennsylvania State University, University of Bern and Moscow State University. He has held several appointments granted by the Academy of Finland including the highly respected Academy Professorship in 2008–2012. Besides being an excellent researcher, Hannu is also an enjoyable teacher. He is a desired speaker at international statistics conferences and a wanted guest lecturer. Hannu takes supervision very seriously, providing excellent guidance and mentoring in all aspects required. Up to date, he has supervised eleven PhD theses—and is still supervising many new promising talents. Interestingly, almost all of Hannu's PhD students have preferred an academic career over a non-academic one. This must have something to do with Hannu's exemplary career and endless positive attitude towards statistics research.

This book consists of 27 contributions written by Hannu's former students, coauthors, colleagues and friends. The book is divided into four parts. Part I starts with some remarks about Hannu's early career, given by his PhD thesis supervisor Professor (emeritus) Elja Arjas, followed by a light introduction to Hannu's publications and coauthors. The remaining three parts of the book cover a wide variety of topics related to Hannu's research interests. In Part II, some recent results in the areas of univariate nonparametric and robust methods are presented. Part III consists of papers concerning modern nonparametric and robust methods in the context of multivariate and functional data. Finally, Part IV is related to Hannu's current research interest on Invariant Coordinate Selection. Also, two contributions on robust methods in signal processing applications are given.

We wish to thank all the authors for their interesting contributions and smooth cooperation during the past 1.5 years. The feedback from the authors has been very positive and we have been pleased to see how every one we asked has been willing to show their gratitude to Hannu via this Festschrift. Our special thanks

go furthermore to those who acted as referees for the contributions. The schedule was occasionally very tight, so without their output, we would not have managed to finish this Festschrift in time. In this context, we would also like to thank Veronika Rosteck from Springer who encouraged us throughout this project and provided help and assistance whenever needed.

Our biggest thanks naturally belong to Hannu who taught us this profession and is still encouraging us on our scientific journey. We hope that his enthusiasm for statistics will continue for a long time and that he will remain an active member of our community.

Tampere, Finland — Klaus Nordhausen
Jyväskylä, Finland — Sara Taskinen
June 2015

Acknowledgements

Contributions 2 to 27 of this festschrift are peer-reviewed. We would like to thank all of the following referees for their excellent work:

Arslan, Olcay	Bing, Li	Chakraborty, Anirvan
Christmann, Andreas	Critchley, Frank	Croux, Christophe
Datta, Somnath	Filzmoser, Peter	Fischer, Daniel
Frahm, Gabriel	Fried, Roland	Ghosh, Anil
Hallin, Marc	Hettmansperger, Tom	Hörmann, Siegfried
Hössjer, Ola	Hubert, Mia	Hunter, David
Ilmonen, Pauliina	Jurečková, Jana	Karvanen, Juha
Kassam, Saleem	Kent, John	Konietschke, Frank
Larocque, Denis	Ley, Christophe	Miettinen, Jari
Möttönen, Jyrki	Nevalainen, Jaakko	Nyblom, Jukka
Ollila, Esa	Paindaveine, Davy	Pawlowsky-Glahn, Vera
Rousseeuw, Peter	Ruiz-Gazen, Anne	Sabolova, Radka
Satten, Glen	Serfling, Robert	Tarr, Garth
Tyler, David	Valkonen, Tuomo	Van Aelst, Stefan
Van Bever, Germain	Verdebout, Thomas	Wied, Dominik
Villa-Vialaneix, Nathalie	Vogel, Daniel	Yohai, Victor
Zamar, Ruben	Zuo, Yijun	

Contents

Contributors

Elja Arjas Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

Giuseppe Bove Dipartimento di Scienze della Formazione, Università degli Studi Roma Tre, Roma, Italy

Delphine Cassart ECARES, Université Libre de Bruxelles, Bruxelles, Belgium

Anirvan Chakraborty Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata, India

Biman Chakraborty School of Mathematics, University of Birmingham, Birmingham, UK

Probal Chaudhuri Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata, India

Patrick D.L. Constable Stony Lodge, Holwell, Sherborne, Dorset, UK

Frank Critchley Department of Mathematics and Statistics, The Open University, Buckinghamshire, UK

Christophe Croux Faculty of Economics and Business, KU Leuven, Leuven, Belgium

Somnath Datta Department of Biostatistics, University of Florida, Gainesville, FL, USA

Susmita Datta Department of Biostatistics, University of Florida, Gainesville, FL, USA

Fikret Er Open Education Faculty, Yunusemre Campus, Anadolu University, Eskisehir, Turkey

Mohamed Fekri Département de Mathématiques, Informatique et Réseaux, Institut National des Postes et Télécommunications, Rabat, Maroc

Peter Filzmoser Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria

Daniel Fischer Natural Resources Institute Finland (Luke), Green Technology, Jokioinen, Finland

School of Health Sciences, University of Tampere, Tampere, Finland

Gabriel Frahm Department of Mathematics/Statistics, Helmut Schmidt University/University of the Federal Armed Forces Germany, Hamburg, Germany

Roland Fried Fakultät Statistik, Technische Universität Dortmund, Dortmund, Germany

Marc Hallin ECARES, Université Libre de Bruxelles, Bruxelles, Belgium

ORFE, Princeton University, Princeton, NJ, USA

Tracey W. Hammel Department of Statistics, Penn State University, University Park, PA, USA

Stefanie Hayoz Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland

Now at Statistics Unit, Swiss Group for Clinical Cancer Research (SAKK), Bern, Switzerland

Thomas P. Hettmansperger Department of Statistics, Penn State University, University Park, PA, USA

Karel Hron Department of Mathematical Analysis and Applications of Mathematics, Palacký University, Olomouc, Czech Republic

Jürg Hüsler Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland

Uwe Jaekel Department of Mathematics and Technology, University of Applied Sciences Koblenz, Remagen, Germany

Jana Jurečková Faculty of Mathematics and Physics, Department of Probability and Statistics, Charles University, Prague 8, Czech Republic

John T. Kent Department of Statistics, University of Leeds, Leeds, UK

Visa Koivunen Department of Signal Processing and Acoustics, Aalto University, Aalto, Finland

Denis Larocque Department of Decision Sciences, HEC Montréal, Montréal, QC, Canada

Denis H.Y. Leung School of Economics, Singapore Management University, Singapore, Singapore

Douglas J. Lorenz Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, USA

Jarmo Lundén Department of Signal Processing and Acoustics, Aalto University, Aalto, Finland

Olusola Samuel Makinde School of Mathematics, University of Birmingham, Birmingham, UK

Jari Miettinen Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

Karl Mosler Institute of Econometrics and Statistics, University of Cologne, Köln, Germany

Jyrki Möttönen Department of Social Research, University of Helsinki, Helsinki, Finland

Jaakko Nevalainen School of Health Sciences, University of Tampere, Tampere, Finland

Klaus Nordhausen Department of Mathematics and Statistics, University of Turku, Turku, Finland

School of Health Sciences, University of Tampere, Tampere, Finland

Jukka Nyblom Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

Viktoria Öllerer Faculty of Economics and Business, KU Leuven, Leuven, Belgium

Esa Ollila Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

Davy Paindaveine ECARES and Department of Mathematics, Université Libre de Bruxelles, Bruxelles, Belgium

Jan Picek Department of Applied Mathematics, Technical University of Liberec, Liberec, Czech Republic

Oleksii Pokotylo Cologne Graduate School, University of Cologne, Köln, Germany

Jing Qin Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, MA, USA

Anne Ruiz-Gazen Toulouse School of Economics, Université Toulouse 1 Capitole, Toulouse, France

Radka Sabolova Department of Mathematics and Statistics, The Open University, Buckinghamshire, UK

Pranab Kumar Sen Departments of Statistics and Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Robert Serfling Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX, USA

Georgy Shevlyakov Institute of Applied Mathematics and Mechanics, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

Mikko J. Sillanpää Department of Mathematical Sciences and Biocenter Oulu, University of Oulu, Oulu, Finland

Pavel Smirnov Institute of Applied Mathematics and Mechanics, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

Ezequiel Smucler Instituto de Cálculo, Universidad de Buenos Aires, Buenos Aires, Argentina

Sara Taskinen Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

David E. Tyler Department of Statistics and Biostatistics, Rutgers – The State University of New Jersey, New Brunswick, NJ, USA

Germain Van Bever Department of Mathematics and Statistics, The Open University, Buckinghamshire, UK

Daniel Vogel Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen, UK

Jue Wang Department of Statistics and Biostatistics, Rutgers – The State University of New Jersey, New Brunswick, NJ, USA

Víctor J. Yohai Departamento de Matemáticas and Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina

Part I
Remarks About Hannu Oja's Career and Publications

Chapter 1
When We Were Very Young: Some Recollections from Hannu Oja's First Years of Academic Life

Elja Arjas

Abstract This chapter is a brief description, written from free memory, of what academic life was like in a very small Finnish university department 40 years ago, and of how some young chaps, themselves still novices—Hannu Oja being one—wanted to make a difference.

It must have been in early fall of 1975 when I met Hannu for the first time. We were sitting in an office of the Mathematics Department of the University of Helsinki, one that I could use during my frequent visits there at the time. It was an informal interview for the position of an Assistant at the newly formed Department of Applied Mathematics and Statistics at the University of Oulu. I had moved to Oulu in January of that same year, and during a relatively short period it appeared possible to expand that small Department a little and advertise some new positions. This was one of them, and its creation had been, as was customary in those days, formally approved by the Parliament as part of the next year's budget of the Finnish government.

I can no longer recall whether there were any other applicants than Hannu. But it was an easy decision for me: Here was a young man with a Master's degree in Statistics from the University of Tampere, who was determined he wanted to pursue post graduate studies in the field, and was willing to move to Oulu for that purpose. At the time of the interview, Hannu had been employed as an actuary at Statistics Finland for some months. But I recall Hannu telling me that he found that job a bit boring, and wanted a change. To me, it sounded like a good recommendation for a job in the free academic world. The rest is history, one may now say. But I would like to use this opportunity to say something about how life and work at Finnish universities looked like at the time, now nearly 40 years ago.

The title "When We Were Very Young" is the title of A.A. Milne's book (1924).

E. Arjas (✉)
Department of Mathematics and Statistics, University of Helsinki, 00014 Helsinki, Finland
e-mail: elja.arjas@helsinki.fi

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_1

It is actually difficult to perceive from today's perspective how homespun many things in the Finnish Academia were then. University of Oulu had been founded in 1959, so it was a fairly new institution in the mid 1970s. Most professors there could be seen as pioneers whose main agenda, in this new territory, was to build up a functioning Department in their respective field, with adequate lectures given to the students. I had myself been appointed to a professorship of applied mathematics and statistics, and as Chair of the Department, with minimal scientific credentials in the former domain and, as I may now say, none in the latter. But I was not an exception. Few people around us seemed to have much experience, or even comprehension, of what it would mean to be actively involved in international level research and publishing. I don't want to say that University of Oulu was somehow "bad" in that respect. The same qualification would have applied to many older institutions, well established in the country. Besides, statistics was not an area which would have had much scientific tradition in Finland, so that there was no clear trace or pattern which we could have followed. And so we started essentially from scratch, trying to make our own thing, as best we could. The key ingredient that made such an attempt at all possible was that there was a lot of talent around, in the junior faculty and students. Being the first university in northern Finland, Oulu University was a major attractor for the many gifted high school students who saw there a route to academic professions, and also up in the social ladder, physically near to their home. Naive as I was, I may have exploited that motivation and drive a little. During the first years in Oulu, I forced all statistics majors to do a compulsory course in measure theoretic probability.

The position of Assistant, to which Hannu was initially hired, was the common junior faculty job in the Finnish universities in those days. It meant a fixed term appointment, usually for 3 years with some possibility for an extension, and the idea was to share one's time, approximately 50–50, between graduate level studies and giving tutorials to students. Thus it was somewhat reminiscent to the American practice of being a PhD student and a TA, but there was a regular monthly salary.

Hannu's involvement in teaching was one of the factors which contributed to making my life as professor easy. I soon learned that Hannu could run the tutorials connected to my lectures quite independently. Early on, I checked that I agreed with his solutions to the assignments which had been given to the students. But since there was never any need for a critical comment, I soon dropped such a practice as redundant. One of my lecture courses was on non-parametric statistics, with Hannu taking care of the tutorials. I had made an attempt to learn something about the topic by reading the textbook by Dickinson-Gibbons in advance. But I suspect that it somehow tuck better for Hannu than for me, as he soon started to entertain his own ideas in the area.

Hannu's first goal in Oulu was to upgrade his Master's degree to that of Licentiate. This academic degree has now practically disappeared, but it was then commonly required before one could move on to actual PhD. It involved both completing formal graduate level courses in some other subject than the major in the Master's degree, and also writing a separate thesis. In Hannu's case it meant studying more mathematics. I should think that such a thorough, and also

time-consuming, preparation was not necessarily a bad idea. Hannu's high school background, from Tampereen klassillinen lyseo, had contained more Latin than mathematics. And my guess, more than factual knowledge, is that the mathematical profile of the statistics curriculum at the University of Tampere was not very high either. Thus such energies for sharpening the mathematics skills were well spent.

The background for Hannu's PhD project was that I had become interested in the concept of stochastic ordering, applying it primarily in reliability and in queuing theory. It became then natural to explore to what extent comparable non-parametric partial orderings, but in some sense of a higher order, could be defined for distribution functions in one dimension. Thus we were led to considering general ways to compare two distributions for their location, spread, skewness, and kurtosis. I still think that the core results in the *Scandinavian Journal of Statistics* (1981) paper, which end up looking at how many times the cumulative distribution functions—after certain transformations—intersect each other, are quite beautiful.

Hannu's PhD thesis had the format of four journal papers and a short summary. I had adopted that bundle thesis format, nippuväitöskirja in Finnish, for my own PhD, and I think it was the first of its kind at least within mathematics in Helsinki. Then rare, but now the dominant format, it gives the student a possibility to go through and learn the various phases required for journal publication, thereby providing important training for the later career as a scientist. It also frees the PhD candidate from writing a separate long thesis, which—after the examination is over—most likely no one will read. But otherwise the requirements were quite stiff: I remember having required that there should be four journal papers, of which at least two should have been accepted for publication. There was also a strong preference for single authorship.

Hannu seemed always very clear on what he wanted to do. On the top of his professional agenda was to carry out independent high calibre research in multivariate statistics. As we know, besides this, Hannu has been productive in biostatistics and signal analysis.

This dedication to research showed sometimes in ways that could be considered nonstandard. For example, after having been promoted to the permanent position of Lecturer in Oulu, Hannu applied for the fixed term position of Senior Assistant, with a somewhat lower salary, because it allowed him more time for his research. But in whatever role Hannu chose to work, the Department of Applied Mathematics and Statistics benefited from his presence enormously. Hannu was the real statistician in the Department. He soon formed his own small research group, supervising students, including two PhDs while still in Oulu.

Wisely enough, the Academy of Finland supported Hannu in this endeavour, both in the form of fixed term research positions and later also by research grants. This support, which has continued to this day, has been crucially important for the success of his work. The Academy is to be congratulated for having correctly identified where seminal work in statistics is carried out in Finland.

Hannu has also been lucky to find talented young people who have been working under his supervision, which has led to an expansion of his research ideas. This nurturing of culture and tradition has greatly enriched the statistical community

country wide. Hannu been the Ambassador of non-parametric statistics, spreading the good word, not only in Oulu, but also in Jyväskylä, Tampere, and now in Turku.

At some point, out of necessity, young people become not-quite-as young. Maybe turning 65 is an appropriate time to notice this. But Hannu does not seem to entertain any idea of slowing down, let alone retirement from research. His engine seems to have the same torque as ever before.

Hannu's motto could well be "Ei tehrä tästä ny numeroo".[1] This is a fitting description, in the local dialect, of the people from Häme, the geographical region around Hannu's hometown Tampere. And it is particularly fitting for Hannu: I can almost hear him saying that.

Given Hannu's extreme level of modesty, I wonder how he feels about celebrating his 65th birthday with this Festschrift? But deeper down, I am sure that he actually understands: Hannu's colleagues and friends want to show their appreciation of his work, and of him as a person. That's not something one could possibly refuse.

[1] It has been popularized by Sinikka Nopola's novel, translated into English with the title *Let's Not Make a Fuss over This*.

Chapter 2
Publication and Coauthorship Networks of Hannu Oja

Daniel Fischer, Klaus Nordhausen, and Sara Taskinen

Abstract In this paper we review Hannu Oja's publications and form coauthor networks based on them. Applying community detection methods to the network formed by all of Hannu's publications shows that his coauthors can be classified into 13 clusters, where two large clusters refer to his methodological research. The network concerning this methodological work is then extended to cover all statistical publications written by Hannu's coauthors. The analysis of the extended network shows that Hannu's coauthors do not form a closed community, but Hannu is involved in many different fields of statistics.

Keywords Bibliography • Community detection • Social network analysis

2.1 Introduction

Hannu (Frans Vilhelm) Oja is one of the most influential statisticians in Finland with an impressive publishing record. In this paper we will briefly review his publications and also have a look at his scientific network. To the benefit of all his friends, colleagues, past and present students as well as the general statistics community, this paper is only a snapshot, as Hannu is still active in research and we are looking forward to many more contributions from him. In this paper we consider those of Hannu's publications that were published before 10 October, 2014. By this date

D. Fischer (✉)
Natural Resources Institute Finland (Luke), Green Technology, 31600 Jokioinen, Finland

School of Health Sciences, University of Tampere, 33014 Tampere, Finland
e-mail: Daniel.Fischer@luke.fi

K. Nordhausen
Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland

School of Health Sciences, University of Tampere, 33014 Tampere, Finland
e-mail: klaus.nordhausen@utu.fi

S. Taskinen
Department of Mathematics and Statistics, University of Jyväskylä, 40014 Jyväskylä, Finland
e-mail: sara.l.taskinen@jyu.fi

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_2

Hannu has published 161 scientific works as listed in Appendix A.1. We will refer to these works using the numbers provided in the Appendix.

Hannu can be regarded as a pioneer in the field of nonparametric methods. Many of his contributions are published in highly ranked statistical journals including *Annals of Statistics, Journal of the American Statistical Association, Journal of the Royal Statistical Society Series A–C, Biometrika*, and *Statistical Science*. Hannu's so far only monograph [131] discusses multivariate methods using spatial signs and ranks. In addition to nonparametric statistics, Hannu has also contributed broadly in the fields of robust multivariate methods, statistical signal processing, and biostatistics. His applied papers fall within the areas of medicine, psychology, and psychiatry. As statistics papers in general receive lower citation numbers, his top three cited papers according to the Web of Science are the applied publications [46], [60], and [66] with 113, 198, and 88 citations, respectively. Of his methodological work, the three most cited papers are [1], [43], and [47] with 77, 43, and 42 citations, respectively. Notice that these figures should be rather taken as lower bounds as the Web of Science does not cover all references.

As Hannu is also interested in making his statistical methodology available to the broad community, he is furthermore involved in nine R packages which are listed in Appendix A.2.

In the following we consider Hannu's role in the scientific community by analyzing the network created by his publications. The following analysis is done in R (R Core Team 2014), using the R packages bibtex (Francois 2014), igraph (Csardi and Nepusz 2006; Kolaczyk and Csardi 2014), and rworldmap (South 2011).

The citation data was extracted from the Web of Science on 10.10.2014 (URL: http://apps.webofknowledge.com) by courtesy of Thomson Reuters.

2.2 The Coauthor Network Created by Hannu Oja's Publications

A basic statistical network is given by a *graph* $G(V, E)$, where V are all N_v *vertices* in the population and the *edges* E between vertices indicate that they are connected in some way. The network can then be described using the so-called *adjacency matrix* $\mathbf{A}$, which is an $N_v \times N_v$ matrix of zeros and ones such that A_{ij} is 1 if vertex V_i is connected to vertex V_j, and otherwise 0. Such networks are usually based on *simple* graphs, which means that the diagonal elements of $\mathbf{A}$ are zero. A network is considered *undirected* when there is no ordering of vertices when considering their connections, which means that $A_{ij} = A_{ji}$ and hence $\mathbf{A}$ is symmetric.

Sometimes only a subset of the whole network is considered. This subset is usually developed around an *ego*-vertex and contains all other vertices next to the ego vertex having a connection to the ego vertex. Such networks are called *egocentric networks*, and the corresponding adjacency matrix is therefore a submatrix of $\mathbf{A}$ of the full graph.

Table 2.1 Collaboration frequencies of Hannu in the coauthor network

Number of joint papers	1	2	3	4	5	6	7	9	10	11	12	14	16	18	19	22
Frequency	77	42	15	6	4	1	2	3	1	1	1	2	1	1	1	1

In this section we consider the coauthor network $G(V, E)$ created by Hannu's publications [1]–[161]. In this network G, the vertices V correspond to Hannu and all his coauthors. Naturally there is then an edge between Hannu and every coauthor, but two distinct coauthors are connected only if they have written at least one joint paper with Hannu. This means that two coauthors might have collaborated sometimes but still might be unconnected in this network if this collaboration does not include Hannu. This network is thus not an egocentric network with ego vertex Hannu out of the network based on all scientific collaborations.

Furthermore, this network is defined as a simple graph meaning that Hannu's 12 single author papers do not contribute to the network. In the remaining 149 publications, Hannu has 159 coauthors (notice that, to the best of our knowledge, we made authors unique by considering, for example, changes in family names). Table 2.1 gives Hannu's collaboration frequencies, which shows that Hannu has nine collaborators with whom he has written ten or more publications.

The number of vertices N_v in this coauthor network is thus 160 and there are altogether $N_e = 666$ edges in G connecting the various vertices.

Figure 2.1 visualizes the network, so that the nodes have different colors depending on a crude and subjective classification into different background fields. The size of the node in this figure depends on the number of publications an author has in this network, and the author's name is printed only if he/she has at least five joint publications with Hannu. To identify the other authors in the network, the list of authors with corresponding label numbers is given in Appendix A.3 (Table 2.2).

Since the network given in Fig. 2.1 includes Hannu's coauthors, the network is naturally a *connected* graph meaning that there is a path from any $v_i \in V$ to $v_j \in V$. However, if Hannu were "removed" from G, the graph would fall apart into ten *components* which would not be connected. The largest component comprises 112 nodes out of 159 coauthors, all the remaining components being relatively small.

The *degree* d_v of a vertex v is the number of edges in E which are connected to v. Having a weight attribute, e.g., the number of joint publications, allows us to multiply the edges by the corresponding weights. The sum of the weighted edges of a vertex is then called *vertex strength*. These quantities can be often used to decide how "central" or "important" a vertex v is in G. However, as this network is neither a "full" network nor an egocentric network, we refrain from doing so here, since the classical measures of network centrality and the measures of centrality for egocentric networks, as described, for example, in Mardsen (2002), would be biased. Further, as Hannu is by design the most central figure, the centrality measure might also be quite uninformative.

Thus, in the following we will focus on searching clusters from our network. In network analysis this is known as *community detection*. We believe that standard

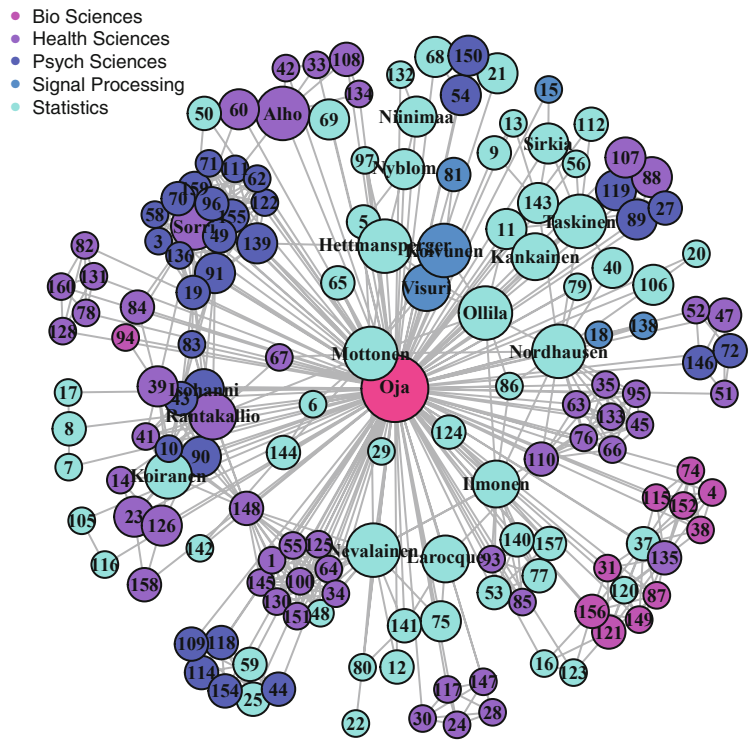

Fig. 2.1 Visualization of the coauthor network of Hannu Oja. *Vertex color* indicates background field and its radius is relative to the number of publications in the network. The name of an author is printed only if the author has at least five joint publications with Hannu

community detection methods can be applied to our network and will not be inherently biased, but reflect in some way the world from Hannu's point of view.

Many methods for community detection have been suggested in the literature (see, for example, Lancichinetti and Fortunato 2009; Fortunato 2010, for an overview), and many of them try to maximize a clustering index called *modularity*. Several variants of modularity have also been introduced, but all of them basically maximize the number of edges within a community, adjusted for the expected number of such edges.

Following Newman (2006), the modularity is defined using the adjacency matrix **A** as

$$Q = \frac{1}{2N_e} \sum_{i=1}^{N_v} \sum_{j=1}^{N_v} \left(A_{ij} - P_{ij}\right) \delta(i,j).$$

In this equation $\delta(i,j) = 1$ if i and j belong to the same community, and otherwise zero. The matrix P_{ij} contains the expected number of edges. To form this matrix,

many null models can be thought of. However, the most common one assumes a random graph having the same degree distribution as the observed graph, i.e., $P_{ij} = (2N_e)^{-1}(\sum_{j=1}^{N_v} A_{ij} \sum_{i=1}^{N_v} A_{ij})$.

In community detection, the trivial solutions, which put all vertices into one cluster or where each vertex forms its own cluster, are usually excluded. Given these constraints, the goal of community detection is then to decide the optimal number of communities and labeling the vertices by maximizing the modularity. Finding the global maximum is, however, for most real networks not tractable computationally, and approximate optimization methods are often used instead. Many of them follow the ideas coming from hierarchical clustering and either successively merge vertices or divide them.

In the following we will motivate the so-called *leading eigenvector method* of Newman (2006), which is, however, not a real hierarchical clustering method.

Assume that we want to divide our network into two communities. Define for this the N_v-vector $\mathbf{s} = (s_1, \ldots, s_{N_v})^\top$, which takes values $+1$ and -1, if the corresponding vector belongs to community 1 or 2, respectively. Notice that $\mathbf{s}^\top \mathbf{s} = N_v$ and $0.5(s_i s_j + 1)$ is 1 if vertices i and j belong to the same group and otherwise 0. Hence we can rewrite the modularity as

$$Q = \frac{1}{4N_e} \sum_{i=1}^{N_v} \sum_{j=1}^{N_v} \left(A_{ij} - P_{ij}\right) (s_i s_j + 1) = \frac{1}{4N_e} \sum_{i=1}^{N_v} \sum_{j=1}^{N_v} \left(A_{ij} - P_{ij}\right) (s_i s_j),$$

where the second equality follows from the fact that $\sum_{i=1}^{N_v} \sum_{j=1}^{N_v} A_{ij} = \sum_{i=1}^{N_v} \sum_{j=1}^{N_v} P_{ij} = 2N_e$.

Hence Q can be defined in matrix notation as

$$Q = \frac{1}{4N_e} \mathbf{s}^\top \mathbf{B} \mathbf{s},$$

where $\mathbf{B}$ is the so-called *modularity matrix*, which is a real and symmetric matrix having elements $B_{ij} = A_{ij} - P_{ij}$.

Let then $\mathbf{B}$ have the eigenvector-eigenvalue decomposition $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$, where $\mathbf{U} = (\mathbf{u}_1, \ldots \mathbf{u}_{N_v})$ is orthogonal and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{N_v})$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{N_v}$. Then

$$Q = \frac{1}{4N_e} \sum_{i=1}^{N_v} \alpha_i^2 \lambda_i,$$

where $\alpha_i = \mathbf{u}_i^\top \mathbf{s}$. Therefore, maximizing Q corresponds to putting as much weight as possible on the largest positive eigenvalues. This would be an easy task if $\mathbf{s}$ would not be restricted to have only entries equal to ± 1. A good approximation seems to

be, however,

$$s_i = \begin{cases} +1 \text{ if } & u_{1i} \geq 0 \\ -1 \text{ if } & u_{1i} < 0, \end{cases}$$

where the magnitude of $|u_{1i}|$ can be interpreted as the strength of belonging to the corresponding community.

Having then two communities, the same method can again be applied to each community, and so on, until no community is divided any further. Many refinements of this methods are available. One can, for example, include also information contained in other eigenvectors, see, for example, Newman (2006) for further information.

Applying this methodology to our graph G and using the number of collaborations to weight the edge degrees as implemented in the igraph package give 18 clusters. As five of these clusters contained only three or fewer authors, we decided to merge them with what we think are their closest communities. The resulting 13 clusters are visualized in Fig. 2.2, where the intercluster and intracluster edges are colored red and black, respectively.

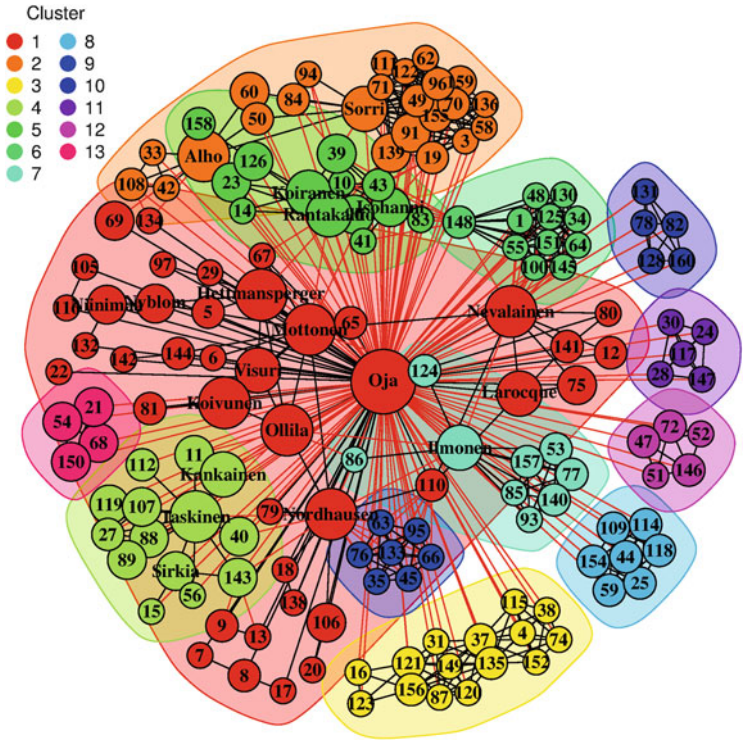

Fig. 2.2 Visualization of the leading eigenvector clustering of network G when using number of joint publications as edge weight

The largest cluster in Fig. 2.2, colored in red and denoted as cluster 1, contains Hannu and 40 coauthors, and forms jointly with cluster 4 the core of his methodological work. Interestingly, cluster 4 mainly contains coauthors from the work conducted during Hannu's stay at the University of Jyväskylä. Hannu's recent side tour to analyzing genetic data, as seen in publications [136,157], forms cluster 3. Clusters 2, 5, and 12 refer to Hannu's applied work conducted during his stay at the University of Oulu. Cluster 2 is formed by the authors of [18,20,25,26,27,32,33,34,37,38]. These publications are related to the large cohort study, which aimed at finding risk factors for acute otitis media. Cluster 5 corresponds to publications [10,11,13,19,23,28,42,66], which originated from the Northern Finland Birth Cohort 1966 (NFBC1966) study and gave insight into how perinatal events affect subsequent morbidity and mortality of children. Cluster 12 refers to publications [35,41].

Some of the other clusters are related to Hannu's different areas of application, many of them referring to only one publication: Cluster 6 corresponds to publication [115], cluster 9 to [121], cluster 10 to [110] and cluster 11 to [99]. In publications [119,144] subjects suffering from schizophrenia were analyzed, and the corresponding authors form cluster 8. Cluster 7 concerns mainly the work Hannu did with his PhD students Pauliina Ilmonen and Sylvia Kiwuwa-Muyingo. Cluster 13 can be considered a more methodological work, since publications [125,126,135] deal with the estimation of population attributable fractions in epidemiological studies.

As can be seen from the above description of the coauthor network, Hannu is a very active scientist with many collaborators. Such collaboration is nowadays highly encouraged, and international contacts are highly emphasized. One motivation for this is that, despite being in the information age, it is assumed that the transfer of knowledge and ideas is faster via direct contact between collaborating researchers (Kretschmer 2004). In order to visualize the grade of internationality of the above network, Fig. 2.3 shows a map of the current locations of Hannu's coauthors to the best of our knowledge.

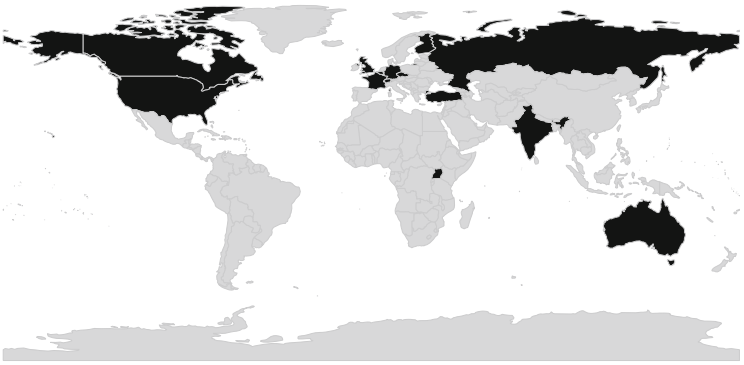

Fig. 2.3 Current location of Hannu Oja's coauthors. The corresponding countries are marked *black*

As seen in Fig. 2.3, Hannu collaborates with researchers around the world, and this has by all means been highly beneficial to the Finnish statistical community. A special group of people who have benefited a lot from Hannu's scientific network are his PhD students. As most of us know, Hannu encourages his students to international collaboration as well as to spending some time abroad during their studies. Without such a widespread scientific network this would not have been possible.

According to the classification visualized in Fig. 2.2, most methodological work seems to be conducted in two clusters. In the following, we work under the assumption that the scientific knowledge and ideas are mainly transferred using personal contacts. We take a closer look at Hannu's network in statistics and also extend the network a bit.

2.3 Extended Statistical Network

To see if Hannu's methodological work is indeed only performed in two communities and also in order to evaluate if these communities can be placed in a larger context, we consider the following network.

First, all Hannu's research articles in the field of mathematics were searched from the Web of Science. Obvious misclassifications were removed by hand. Then all coauthors of the remaining 67 publications were extracted, and their publications with the same restrictions (mathematics & articles) were searched. The publication period under consideration was from 1981 to 2014. It should be noted that our main search term, author name plus initial, of course leads to many false positives. Some obvious mistakes were again removed by hand, but the network must be considered with some reservation, as besides the false positives, the Web of Science also does not index all publications for all authors.

The resulting network consists of 946 papers written by 653 authors, out of which 98 are single author papers. This extended network has 1861 edges and is visualized in Fig. 2.4. In this figure, the vertex size is again relative to the number of publications of the corresponding author. For visualization purposes, we only print the names of the authors with at least 15 publications as well as of those responsible for this paper.

The figure clearly indicates that Hannu's coauthors do not form a closed community but there are many "ways out" for Hannu's knowledge and ideas to conquer the world.

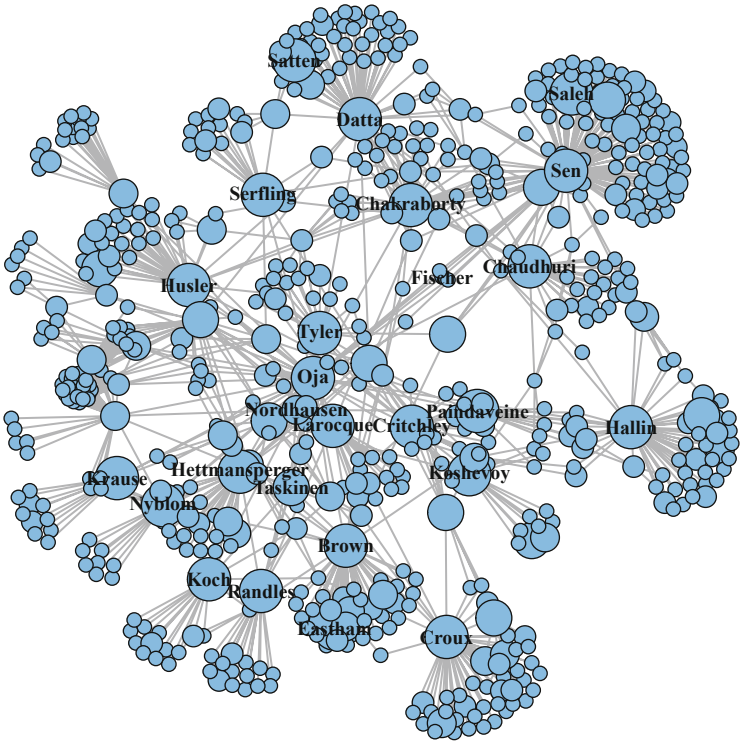

Fig. 2.4 Visualization of extended network. Vertex size reflects the amount a researcher published and labels are only given for authors with at least 15 publications and for the authors of this paper

Like in the analysis of the previous network, we again refrain from providing centrality measures as Hannu is by design the most central vertex and the network is again neither a standard network nor an egocentric network. Hence we again focus on clustering the network. As the network is larger than before, we take the greedy hierarchical clustering method implemented in igraph, where the edges were again weighted by the number of joint publications. The resulting 14 clusters are visualized in Fig. 2.5. Also this figure seems to support the idea that Hannu is involved in many different communities, which is easily understandable considering his various interests in statistics.

It would be nice to extend this network until the whole statistics field would be covered in order to see how central a player Hannu is in that network (Hannu's Erdös number is 3, for example.). It would also be nice to look at the dynamic structure of this network in order to see how Hannu's coauthorship network develops and changes over time. However, we decided to postpone such an analysis until Hannu decides to stop publishing scientific articles, which we know will not happen in the near future. We are sure that Hannu has still many valuable contributions to the field of statistics and will also start many new fruitful collaborations!

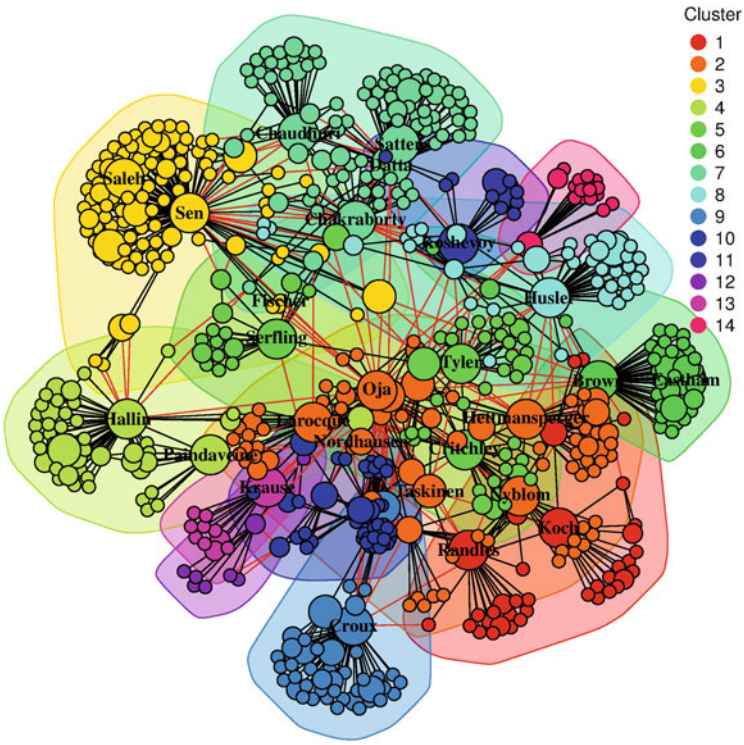

Fig. 2.5 Visualization of hierarchical clustering of the extended network when using number of joint publications as edge weight

Acknowledgements The authors are grateful to Markus Mattilainen and Joni Virta for helping to extract the data from Web of Science. And for the comments given by Nathalie Villa-Vialaneix and David Hunter which helped us to improve the paper tremendously. Certain data included herein are derived from the Web of Science® prepared by Thomson Reuters Scientific, Inc. (Thomson®), Philadelphia, Pennsylvania, USA:©Copyright THOMSON REUTERS® 2014. All rights reserved. This work was supported by the Academy of Finland under Grants 251965 and 268703.

Appendices

A.1 Publications of Hannu Oja

[1] Oja, H.: On location, scale, skewness and kurtosis of univariate distributions. Scand. J. Stat. **8**, 154–168 (1981)

[2] Oja, H.: Two location and scale-free goodness-of-fit tests. Biometrika **68**, 637–640 (1981)

[3] Oja, H.: Descriptive statistics for multivariate distributions. Stat. Probab. Lett. **1**, 327–332 (1983)

[4] Oja, H.: New tests for normality. Biometrika **70**, 297–299 (1983)

[5] Oja, H.: Partial ordering of distributions. In: Kotz, S., Johnson, N.L. (eds.) Encyclopedia of Statistical Science, pp. 490–494. Wiley, New York (1985)

[6] Oja, H., Niinimaa, A.: Asymptotical properties of the generalized median in the case of multivariate normality. J. R. Stat. Soc. Ser. B **47**, 372–377 (1985)

[7] Oja, H.: On permutation tests in multiple regression and analysis of covariance problems. Aust. J. Stat. **29**, 81–100 (1987)

[8] Oja, H., Nyblom, J.: On bivariate sign tests. J. Am. Stat. Assoc. **84**, 249–259 (1989)

[9] Niinimaa, A., Oja, H., Tableman, M.: The finite-sample breakdown point of the Oja bivariate median and of the corresponding half-samples version. Stat. Probab. Lett. **10**, 325–328 (1990)

[10] Rantakallio, P., Hartikainen-Sorri, A.-L., Sipilä, P., Oja, H., ElSaid, U., Koiranen, M.: Computer-based perinatal risk prediction in a geographically defined parturient population - an intervention study. Pediatr. Res. **28**, 305 (1990)

[11] Rantakallio, P., Oja, H.: Perinatal risk for infants of unmarried mothers during 20 years. Early Hum. Dev. **22**, 157–169 (1990)

[12] Oja, H., Koiranen, M., Rantakallio, P.: Fitting mixture models to birth weight data: a case study. Biometrics **47**, 883–897 (1991)

[13] Rantakallio, P., Oja, H., Koiranen, M.: Has intrauterine weight gain curve changed in shape? Paediatr. Perinat. Epidemiol. **5**, 201–220 (1991)

[14] Brown, B.M., Hettmansperger, T.P., Nyblom, J., Oja, H.: On certain bivariate sign tests and medians. J. Am. Stat. Assoc. **87**, 127–135 (1992)

[15] Hettmansperger, T.P., Nyblom, J., Oja, H.: On multivariate notions of sign and rank. In: Dodge, Y. (ed.) L_1-Statistical Analysis and Related Methods, pp. 267–278. Elsevier/North Holland, Amsterdam (1992)

[16] Niinimaa, A., Oja, H., Nyblom, J.: Algorithm AS 277: The Oja bivariate median. Appl. Stat. **41**, 611–617 (1992)

[17] Sipilä, P., Hartikainen-Sorri, A.-L., Oja, H., von Wendt, L.: Perinatal outcome of pregnancies complicated by vaginal bleeding. Br. J. Obstet. Gynaecol. **99**, 959–963 (1992)

[18] Alho, O.P., Kilkku, O., Oja, H., Koivu, M., Sorri, M.: Control of the temporal aspect when considering risk factors for acute otitis media. Arch. Otolaryngol. Head Neck Surg. **119**, 444–449 (1993)

[19] Isohanni, M., Oja, H., Rantakallio, P., Moilanen, I., Koiranen, M.: The relation of the teenage smoking and alcohol use, with special reference to the deficient family background. Scand. J. Soc. Med. **21**, 24–30 (1993)

[20] Alho, O.P., Koivu, M., Sorri, M., Oja, H., Kilkku,O.: Which children are being operated on for recurrent acute otitis media? Arch. Otolaryngol. Head Neck Surg. **120**, 807–811 (1994)

[21] Hettmansperger, T.P., Nyblom, J., Oja, H.: Affine invariant multivariate one-sample sign test. J. R. Stat. Soc. Ser. B **56**, 221–234 (1994)

[22] Hettmansperger, T.P., Oja, H.: Affine invariant multivariate multisample sign tests. J. R. Stat. Soc. Ser. B **56**, 235–249 (1994)

[23] Isohanni, M., Oja, H., Moilanen, I., Koiranen, M.: Teenage alcohol drinking and nonstandard family background. Soc. Sci. Med. **38**, 1565–1574 (1994)

[24] Sipilä, P., Hartikainen, A.-L., von Wendt, L., Oja, H.: Changes in risk factors for unfavourable pregnancy outcome among singletons during 20 years. Acta Obstet. Gynecol. Scand. **73**, 612–618 (1994)
[25] Alho, O., Jokinen, K., Pirilä, T., Ilo, A., Oja, H.: Acute epiglottitis and infant conjugate haemophilus influenzae type b vaccination in northern Finland. Arch. Otolaryngol. Head Neck Surg. **121**, 898–902 (1995)
[26] Alho, O.P., Oja, H., Koivu, M., Sorri, M.: Chronic otitis media with effusion in infancy - how frequent is it and how does it develop? Arch. Otolaryngol. Head Neck Surg. **121**, 432–436 (1995)
[27] Alho, O.P., Oja, H., Koivu, M., Sorri, M.: Risk factors for chronic otitis media with effusion in infancy. each acute otitis media episode induces a high but transient risk. Arch. Otolaryngol. Head Neck Surg. **121**, 839–843 (1995)
[28] Isohanni, K., Oja, H., Moilanen, I., Koiranen, M., Rantakallio, P.: Smoking or quitting during pregnancy: associations with background and future social factors. Scand. J. Soc. Med. **23**, 32–38 (1995)
[29] Möttönen, J., Oja, H.: Multivariate spatial sign and rank methods. J. Nonparametr. Stat. **5**, 201–213 (1995)
[30] Möttönen, J., Oja, H., Krause, U., Rantakallio, P.: Application of random coefficient regression model to myopia data: a case study. Biom. J. **37**, 657–672 (1995)
[31] Niinimaa, A., Oja, H.: On the influence functions of certain bivariate medians. J. R. Stat. Soc. Ser. B **57**, 565–574 (1995)
[32] Alho, O.P., Läärä, E., Oja, H.: How should relative risk estimates for acute otitis media in children aged less than 2 years be perceived? J. Clin. Epidemiol. **49**, 9–14 (1996)
[33] Alho, O.P., Läärä, E., Oja, H.: Public health impact of various risk factors for acute otitis media in northern Finland. Am. J. Epidemiol. **143**, 1149–1156 (1996)
[34] Alho, O.P., Läärä, E., Oja, H.: What is the natural history of recurrent acute otitis media in infancy? J. Fam. Pract. **43**, 258–264 (1996)
[35] Kiuttu, J., Larivaara, P., Väisänen, E., Keinänen-Kiukaanniemi, S., Oja, H.: The effects of family systems medicine training on the practice orientations of general practitioners. Fam. Syst. Health **14**, 453–462 (1996)
[36] Luukkonen, S., Koivunen, V., Oja, H.: Multivariate median filter for correlated noise. In: Proceedings of the 1996 IEEE Nordic Signal Processing Symposium, pp. 411–414 (1996)
[37] Oja, H., Alho, O.P., Läärä, E.: Model-based estimation of the excess fraction (attributable fraction): day care and middle ear infection. Stat. Med. **15**, 1519–1534 (1996)
[38] Sorri, M., Alho, O.P., Oja, H.: Dynamic multivariate modelling: day care and consultation rate for acute otitis media. Acta Otolaryngol. **116**, 299–301 (1996)
[39] Brown, B.M., Hettmansperger, T.P., Möttönen, J., Oja, H.: Rank plots in the affine invariant case. In: Dodge, Y. (ed.) L_1-Statistical Procedures and Related Topics: Papers from the 3rd International Conference on L_1-Norm and Related Methods held in Neuchâtel, 11–15 August 1997, pp. 351–362 (1997)
[40] Hettmansperger, T.P., Möttönen, J., Oja, H.: Affine invariant multivariate one-sample signed-rank tests. J. Am. Stat. Assoc. **92**, 1591–1600 (1997)
[41] Kiultu, J., Väisänen, E., Larivaara, P., Keinänen-Kiukaanniemi, S., Oja, H.: Family systems medicine training: helping to meet psychiatric and psychosomatic problems. Nord. J. Psychiatry **51**, 259–265 (1997)
[42] Mäkikyrö, T., Isohanni, M., Moring, J., Oja, H., Hakko, H., Jones, P., Rantakallio, P.: Is a child's risk of early onset schizophrenia increased in the highest social class? Schizophr. Res. **23**, 245–252 (1997)
[43] Möttönen, J., Oja, H., Tienari, J.: On the efficiency of multivariate spatial sign and rank tests. Ann. Stat. **25**, 542–552 (1997)
[44] Pirilä, T., Talvisara, A., Alho, O.P., Oja, H.: Physiological fluctuations in nasal resistance may interfere with nasal monitoring in the nasal provocation test. Acta Otolaryngol. **117**, 596–600 (1997)

[45] Sorri, M., Mäki-Torkko, E., Järvelin, M.-R., Oja, H.: Prevalence figures for mild hearing impairments cannot be based on clinical data. Acta Otolaryngolog. **117**, 179–181 (1997)
[46] Wahlberg, K.-E., Wynne, L.C., Oja, H., Keskitalo, P., Pykäläinen, L., Lahti, I., Moring, J., Naarala, M., Sorri, A., Seitamaa, M., Läksy, K., Kolassa, J., Tienari, P.: Gene-environment interaction in vulnerability to schizophrenia: Findings from the Finnish adoptive family study of schizophrenia. Am. J. Psychiatry **154**,355–362 (1997)
[47] Chakraborty, B., Chaudhuri, P., Oja, H.: Operating transformation retransformation on spatial median and angle test. Stat. Sin. **8**, 767–784 (1998)
[48] Hettmansperger, T. P., Möttönen, J., Oja, H.: Affine invariant multivariate rank tests for several samples. Stat. Sin. **8**, 785–800 (1998)
[49] Koivunen, V., Luukkonen, S., Oja, H.: Affine equivariance in multichannel OS-filtering. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 5, pp. 2881–2884 (1998)
[50] Mäki-Torkko, E.M., Järvelin, M.-R., Sorri, M.J., Muhli, A.A., Oja, H.: Aetiology and risk indicators of hearing impairments in a 1-year birth cohort for 1985–86 in northern Finland. Scand. Audiol. **27**, 237–247 (1998)
[51] Möttönen, J., Hettmansperger, T.P., Oja, H., Tienari, J.: On the efficiency of affine invariant multivariate rank tests. J. Multivar. Anal. **66**, 118–132 (1998)
[52] Hettmansperger, T.P., Möttönen, J., Oja, H.: The geometry of the affine invariant multivariate sign and rank methods. J. Nonparametr. Stat. **11**, 271–285 (1999)
[53] Hettmansperger, T.P., Oja, H., Visuri, S.: Discussion of "Multivariate analysis by data depth: descriptive statistics, graphics and inference" by Liu, Parelius and Singh. Ann. Stat. **3**, 845–853 (1999)
[54] Möttönen, J., Koivunen, V., Oja, H.: Robust autocovariance estimation based on sign and rank correlation coefficients.In: Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics, pp. 187–190(1999)
[55] Niinimaa, A., Oja, H.: Multivariate median. In: Kotz, S., Read, C.B., Banks, D. (eds.), Encyclopedia of Statistical Sciences, vol. 3, pp. 497–505. Wiley, New York (1999)
[56] Oja, H.: Affine invariant multivariate sign and rank tests and corresponding estimates: a review. Scand. J. Stat. **26**, 319–343 (1999)
[57] Visuri, S., Koivunen, V., Möttönen, J., Oja, H.: Matrix perturbations in covariance and autocovariance matrix estimators. In: Proceedings of the 1999 Finnish Signal Processing Symposium, pp. 142–146 (1999)
[58] Visuri, S., Oja, H., Koivunen, V.: Multichannel signal processing using spatial rank covariance matrices. In: Proceedings of the IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP'99), pp. 75–79 (1999)
[59] Sandberg, S., Paton, J.Y., Ahola, S., McCann, D.C., McGuinness, D., Hillary, C.R., Oja, H.: Stressi lisää lasten astmakohtausten riskiä. Duodemic **116**, 2305–2306 (2000)
[60] Sandberg, S., Paton, J.Y., Ahola, S., McCann, D.C., McGuinness, D., Hillary, C.R., Oja, H.: The role of acute and chronic stress in asthma attacks in children. Lancet **356**, 982–988 (2000)
[61] Visuri, S., Koivunen, V., Oja, H.: Sign and rank covariance matrices. J. Stat. Plan. Inference **91**, 557–575 (2000)
[62] Visuri, S., Oja, H., Koivunen, V.: Nonparametric method for subspace based frequency estimation. In: Proceedings of 10th European Signal Processing Conference 2000 (EUSIPCO 2000), pp. 1261–1264 (2000)
[63] Visuri, S., Oja, H., Koivunen, V.: Nonparametric statistics for DOA estimation in the presence of multipath. In: Proceedings of the First IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM 2000), pp. 356–360 (2000)
[64] Visuri, S., Oja, H., Koivunen, V.: Robust subspace DOA estimation for wireless communications. In: The IEEE Annual Vehicular Technology Conference VTC2000, pp. 2551–2555 (2000)
[65] Wahlberg, K.E., Wynne, L.C., Oja, H., Keskitalo, P., Anais-Tanner, H., Koistinen, P., Tarvainen, T., Hakko, H., Lahti, I., Moring, J., Naarala, M., Sorri, A., Tienari, P. Thought

disorder index of Finnish adoptees and communication deviance of their adoptive parents. Psychol. Med. **30**, 127–136 (2000)

[66] Isohanni, M., Jones, P.B., Moilanen, K., Rantakallio, P., Veijola, J., Oja, H., Koiranen, M., Jokelainen, J., Croudace, T., Järvelin, M.: Early developmental milestones in adult schizophrenia and other psychoses. A 31-year follow-up of the northern Finland 1966 birth cohort. Schizophr. Res. **52**, 1–19 (2001)

[67] Sandberg, S., McCann, D.C., Ahola, S., Oja, H., Paton, J.Y., McGuinness, D.: Positive experiences and the relationship between stress and asthma in children. Acta Paediatr. **91**, 152–158 (2001)

[68] Visuri, S., Oja, H., Koivunen, V.,: Blind channel identification using robust subspace estimation. In: Proceedings of the 11th IEEE Signal Processing Workshop on Statistical Signal Processing, pp. 281–284 (2001)

[69] Visuri, S., Oja, H., Koivunen, V.: Subspace-based direction of arrival estimation using nonparametric statistics. In: IEEE Transactions on Signal Processing, vol. 49, pp. 2060–2073 (2001)

[70] Croux, C., Ollila, E., Oja, H.: Sign and rank covariance matrices: statistical properties and application to principal component analysis. In: Dodge, Y. (ed.) Statistical Data Analysis Based on the L_1-Norm and Related Methods, pp. 257–269. Birkhäuser Verlag, Basel (2002)

[71] Ollila, E., Hettmansperger, T.P., Oja, H.: Estimates of regression coefficients based on sign covariance matrix. J. R. Stat. Soc. Ser. B **64**, 447–466 (2002)

[72] Ronkainen, T., Oja, H., Orponen, P.: Computation of the multivariate Oja median. In: Dutter, R., Filzmoser, P., Gather, U., Rousseeuw, P.J. (eds.), Developments in Robust Statistics, pp. 344–359. Springer, Heidelberg (2002)

[73] Taskinen, S., Kankainen, A., Oja, H.: Tests of independence based on sign and rank covariances. In: Dutter, R., Filzmoser, P., Gather, U., Rousseeuw, P.J. (eds.), Developments in Robust Statistics, pp. 387–403. Springer, Heidelberg (2002)

[74] Koshevoy, G.A., Möttönen, J., Oja, H.: A scatter matrix estimate based on the zonotope. Ann. Stat. **31**, 1439–1459 (2003)

[75] Möttönen, J., Hüsler, J., Oja, H.: Multivariate nonparametric tests in a randomized complete block design. J. Multivar. Anal. **85**, 106–129 (2003)

[76] Nadar, M., Hettmansperger, T.P., Oja, H.: The asymptotic covariance matrix of the Oja median. Stat. Probab. Lett. **64**, 431–442 (2003)

[77] Oja, H.: Multivariate M-estimates of location and shape. In: Höglund, R., Jäntti, M., Rosenqvist, G. (eds.), Statistics, Econometrics and Society. Essays in Honor of Leif Nordberg. Statistics Finland (2003)

[78] Ollila, E., Oja, H., Croux, C.: The affine equivariant sign covariance matrix: asymptotic behavior and efficiencies. J. Multivar. Anal. **87**, 328–355 (2003)

[79] Ollila, E., Oja, H., Koivunen, V.: Estimates of regression coefficients based on lift rank covariance matrix. J. Am. Stat. Assoc. **99**(461), 90–98 (2003)

[80] Taskinen, S., Kankainen, A., Oja, H.: Sign test of independence between two random vectors. Stat. Probab. Lett. **62**, 9–21 (2003)

[81] Taskinen, S., Kankainen, A., Oja, H.: Tests of independence based on sign and rank covariances. In: Dutter, R., Filzmoser, P., Gather, U., Rousseeuw, P.J. (eds.), Developments in Robust Statistics, pp. 387–403. Physica-Verlag, HD (2003)

[82] Topchii, A., Tyurin, Y., Oja, H.: Inference based on the affine invariant multivariate Mann–Whitney–Wilcoxon statistic. J. Nonparametr. Stat. **15**, 403–419 (2003)

[83] Visuri, S., Ollila, E., Koivunen, v., Möttönen, J., Oja, H.: Affine equivariant multivariate rank methods. J. Stat. Plan. Inference **114**, 161–185 (2003)

[84] Kankainen, A., Taskinen, S., Oja, H.: On Mardia's tests of multinormality. In: Hubert, M., Pison, G., Stryuf, A., Van Aelst, S. (eds.), Statistics for Industry and Technology, pp. 152–164. Birkhäuser, Basel (2004)

[85] Koshevoy, G.A., Möttönen, J., Oja, H.: On the geometry of multivariate L_1 objective functions. Allg. Stat. Arch. **88**, 137–154 (2004)

[86] Möttönen, J., Oja, H., Serfling, R.J.: Multivariate generalized spatial signed-rank methods. J. Stat. Res. **39**, 19–42 (2004)
[87] Oja, H., Randles, R.H.: Multivariate nonparametric tests. Stat. Sci. **19**, 598–605 (2004)
[88] Ollila, E., Croux, C., Oja, H.: Influence function and asymptotic efficiency of the affine equivariant rank covariance matrix. Stat. Sin. **14**(1), 297–316 (2004)
[89] Taskinen, S., Kankainen, A., Oja, H.: Rank scores tests of multivariate independence. In: Hubert, M., Pison, G., Stryuf, A., Van Aelst, S. (eds.), Statistics for Industry and Technology, pp. 329–341. Birkhäuser, Basel (2004)
[90] Oja, H.: Discussion of "Breakdown and groups" by P. L. Davies and U. Gather. Ann. Stat. **33**(3), 1000–1004 (2005)
[91] Oja, H., Paindaveine, D.: Optimal signed-rank tests based on hyperplanes. J. Stat. Plan. Inference **135**, 300–323 (2005)
[92] Taskinen, S., Oja, H., Randles, R.H.: Multivariate nonparametric tests of independence. J. Am. Stat. Assoc. **100**, 916–925 (2005)
[93] Busarova, D., Tyurin, Y., Möttönen, Y., Oja, H.: Multivariate Theil estimator with the corresponding test. Math. Methods Stat. **15**, 1–19 (2006)
[94] Hallin, M., Oja, H., Paindaveine, D.: Semiparametrically efficient rank-based inference for shape. II. Optimal R-estimation of shape. Ann. Stat. **34**, 2757–2789 (2006)
[95] Hartikainen, A., Oja, H.: On nonparametric discrimination rules. In: Liu, R. (ed.), Data Depth: Robust Multivariate Analysis, Computational Geometry, and Applications. AMS DIMACS Book Series, pp. 61–70 (2006)
[96] Nevalainen, J., Oja, H.: SAS/IML macros for a multivariate analysis of variance based on spatial signs. J. Stat. Softw. **16**, 1–17 (2006)
[97] Nordhausen, K., Oja, H., Tyler, D.E.: On the efficiency of invariant multivariate sign and rank test. In: Liski, E.P., Isotalo, J., Niemelä, J., Puntanen, S., Styan, G.P.H. (eds.), Festschrift for Tarmo Pukkila on his 60th Birthday, pp. 217–231. University of Tampere, Tampere (2006)
[98] Oja, H., Sirkiä, S., Eriksson, J.: Scatter matrices and independent component analysis. Aust. J. Stat. **35**, 175–189 (2006)
[99] Saastamoinen, A., Oja, H., Huupponen, E., Värri, A., Hasan, J., Himanen, S.L.: Topographic differences in mean computational sleep depth between healthy controls and obstructive sleep apnoea patients. J. Neurosci. Methods **157**(1), 178–184 (2006)
[100] Taskinen, S., Croux, C., Kankainen, A., Ollila, E., Oja, H.: Influence functions and efficiencies of the canonical correlation and vector estimates based on scatter and shape matrices. J. Multivar. Anal. **97**, 359–384 (2006)
[101] Kankainen, A., Taskinen, S., Oja, H.: Tests of multinormality based on location vectors and scatter matrices. Stat. Methods Appl. **16**, 357–379 (2007)
[102] Larocque, D., Nevalainen, J., Oja, H.: A weighted multivariate sign test for cluster correlated data. Biometrika **94**, 267–283 (2007)
[103] Nevalainen, J., Larocque, D., Oja, H.: On the multivariate spatial median for clustered data. Can. J. Stat. **35**, 215–283 (2007)
[104] Nevalainen, J., Larocque, D., Oja, H.: A weighted spatial median for clustered data. Stat. Methods Appl. **15**, 355–379 (2007)
[105] Nordhausen, K., Oja, H., Ollila, E.: Robust ICA based on two scatter matrices. In: Aivazian, S., Filzmoser, P., Kharin, Y. (eds.) Proceedings of the 8th International Conference on Computer Data Analysis and Modeling, pp. 84–91. Minsk Publishing Center BSU, Minsk (2007)
[106] Oja, H., Critchley, F.: Discussion of the paper "A survey on robust statistics" by S. Morgenthaler. Stat. Methods Appl. **15**, 271–293 (2007)
[107] Sirkiä, S., Taskinen, S., Oja, H.: Symmetrised M-estimators of scatter. J. Multivar. Anal. **98**, 1611–1629 (2007)
[108] Taskinen, S., Sirkiä, S., Oja, H.: Independent component analysis based on symmetrised scatter estimators. Comput. Stat. Data Anal. **51**, 5103–5111 (2007)

[109] Larocque, D., Nevalainen, J., Oja, H.: One-sample location tests for multilevel data. J. Stat. Plan. Inference **138**, 2469–2482 (2008)
[110] Lindroos, B., Mäenpää, K., Ylikomi, T., Oja, H., Suuronen, R., Miettinen, S.: Characterisation of human dental stem cells and buccal mucosa fibroblasts. Biochem. Biophys. Res. Commun. **368**(2), 329–335 (2008)
[111] Nevalainen, J., Möttönen, J., Oja, H.: A spatial rank test and corresponding estimators for several samples. Stat. Probab. Lett. **78**, 661–668 (2008)
[112] Nordhausen, K., Oja, H., Ollila, E.: Robust independent component analysis based on two scatter matrices. Aust. J. Stat. **37**, 91–100 (2008)
[113] Nordhausen, K., Oja, H., Tyler, D.E.: Tools for exploring multivariate data: The package ICS. J. Stat. Softw. **28**(6) , 1–31 (2008)
[114] Ollila, E., Oja, H., Koivunen, V.: Complex-valued ICA based on a pair of generalized covariance matrices. Comput. Stat. Data Anal. **52**, 3789–3805 (2008)
[115] Uusitalo, L., Nevalainen, J., Niinistö, S., Alfthan, G., Sundvall, J., Korhonen, T., Kenward, M.G., Oja, H., Veijola, R., Simell, O., Ilonen, J., Knip, M., Virtanen, S.M.: Serum alpha- and gamma-tocopherol concentrations and risk of advanced beta cell autoimmunity in children with HLA-conferred susceptibility to type 1 diabetes mellitus. Diabetologia **51**, 773–780 (2008)
[116] Chaudhuri, P., Ghosh, A.K., Oja, H.: Classification based on hybridization of parametric and nonparametric classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **31**(7), 1153–1164 (2009)
[117] Haataja, R., Larocque, D., Nevalainen, J., Oja, H.: A weighted multivariate signed-rank test for cluster-correlated data. J. Multivar. Anal. **100**, 1107–1119 (2009)
[118] Nordhausen, K., Oja, H., Paindaveine, D.: Signed-rank tests for location in the symmetric independent component model. J. Multivar. Anal. **100**, 821–834 (2009)
[119] Rantanen, H., Koivisto, A.M., Salokangas, R.K., Helminen, M., Oja, H., Pirkola, S., Wahlbeck, K., Joukamaa, M.: Five-year mortality of Finnish schizophrenia patients in the era of deinstitutionalization. Soc. Psychiatry Psychiatr. Epidemiol. **44**, 135–142 (2009)
[120] Sirkiä, S., Taskinen, S., Oja, H., Tyler, D.E.: Tests and estimates for shape based on spatial signs and ranks. J. Nonparametr. Stat. **21**, 155–176 (2009)
[121] Tahvanainen, A., Leskinen, M., Koskela, J., Ilveskoski, E., Nordhausen, K., Oja, H., Kähönen, M., Kööbi, T., Mustonen, J., Pörsti, I.: Ageing and cardiovascular responses to head-up tilt in healthy subjects. Atherosclerosis **2007**(2), 445–451 (2009)
[122] Tyler, D.E., Critchley, F., Dümbgen, L., Oja, H.: Invariant co-ordinate selection. J. R. Stat. Soc. Ser. B **71**, 549–592 (2009)
[123] Ilmonen, P., Nevalainen, J., Oja, H.: Characteristics of multivariate distributions and the invariant coordinate system. Stat. Probab. Lett. **80**, 1844–1853 (2010)
[124] Ilmonen, P., Nordhausen, K., Oja, H., Ollila, E.: A new performance index for ICA: Properties, computation and asymptotic analysis. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) Latent Variable Analysis and Signal Separation - 9th International Conference, LVA/ICA 2010, St. Malo, France, 27–30 September 2010. Proceedings of Lecture Notes in Computer Science, vol. 6365, pp. 229–236. Springer (2010)
[125] Laaksonen, M.A., Knekt, P., Härkänen, T., Virtala, E., Oja, H.: Estimation of the population attributable fraction for mortality in a cohort study using a piecewise constant hazards model. Am. J. Epidemiol. **171**, 837–847 (2010)
[126] Laaksonen, M.A., Härkänen, T., Knekt, P., Virtala, E., Oja, H.: Estimation of population attributable fraction (PAF) for disease occurrence in a cohort study design. Stat. Med. **29**(7–8), 860–874 (2010)
[127] Larocque, D., Haataja, R., Nevalainen, J., Oja, H.: Two sample tests for the nonparametric Behrens-Fisher problem with clustered data. J. Nonparametric Stat. **22**, 755–771 (2010)
[128] Möttönen, J., Nordhausen, K., Oja, H.: Asymptotic theory of the spatial median. In: Antoch, J., Huskova, M., Sen, P.K. (eds.), Nonparametrics and Robustness in Modern Statistical

Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková, pp. 182–193. Institute of Mathematical Statistics, Beachwood, OH (2010)
[129] Nevalainen, J., Larocque, D., Oja, H., Pörsti, I.: Nonparametric analysis of multivariate clustered data. J. Am. Stat. Assoc. **105**, 864–872 (2010)
[130] Nordhausen, K., Oja, H.: Three scatter matrices and independent subspace analysis. In: Aivazian, S., Filzmoser, P., Kharin, Y. (eds.) Proceedings of the 9th International Conference on Computer Data Analysis and Modeling, pp. 93–100. Minsk Publishing Center BSU, Mink (2010)
[131] Oja, H.: Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks. Springer (2010)
[132] Rauhala, H.E., Jalava, S.E., Isotalo, J., Bracken, H., Lehmusvaara, S., Tammela, T.L., Oja, H., Visakorpi, T.: mir-193b is an epigenetically regulated putative tumor suppressor in prostate cancer. Int. J. Cancer **127**(6), 1363–1372 (2010)
[133] Taskinen, S., Sirkiä, S., Oja, H.: *k*-step shape estimators based on spatial signs and ranks. J. Stat. Plan. Inference **140**(11), 3376–3388 (2010)
[134] Kiwuwa-Muyingo, S., Oja, H., Walker, A.S., Ilmonen, P., Levin, J., Todd, J.: Clustering based on adherence data. Epidemiol. Perspect. Innov. (Methodology) **8**(3) (2011)
[135] Laaksonen, M.A., Virtala, E., Knekt, P., Oja, H., Härkänen, T.: SAS macros for calculation of population attributable fraction in a cohort study design. J. Stat. Softw. **43**(7), 1–25 (2011)
[136] Mattila, H., Schindler, M., Isotalo, J., Ikonen, T., Vihinen, M., Oja, H., Tammela, T.L.J., Wahlfors, T., Schleutker, J.: NMD and microRNA expression profiling of the HPCX1 locus reveal MAGEC1 as a candidate prostate cancer predisposition gene. BMC Cancer 2011 **11**(327) (2011)
[137] Nordhausen, K., Ilmonen, P., Mandal, A., Oja, H., Ollila, E.: Deflation-based FastICA reloaded. In: Proceedings of 19th European Signal Processing Conference 2011 (EUSIPCO 2011), pp. 1854–1858 (2011)
[138] Nordhausen, K., Oja, H.: Discussion on the paper "The asymptotic efficiency of the spatial median for elliptically symmetric distributions" by Andrew Magyar and D.E. Tyler. Sankhya Ser. B **73**, 188–191 (2011)
[139] Nordhausen, K., Oja, H.: Independent subspace analysis using three scatter matrices. Aust. J. Stat. **40**(1&2), 93–101 (2011)
[140] Nordhausen, K., Oja, H.: Multivariate L_1 methods: the package MNM. J. Stat. Softw. **43**, 1–28 (2011)
[141] Nordhausen, K., Oja, H.: Scatter matrices with independent block property and ISA. In: Proceedings of 19th European Signal Processing Conference 2011 (EUSIPCO 2011), pp. 1738–1742 (2011)
[142] Nordhausen, K., Oja, H., Ollila, E.: Multivariate models and the first four moments. In: Hunter, D.R., Richards, D.St.P., Rosenberger, J.L. (eds.) Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P Hettmansperger, pp. 267–287. World Scientific, Singapore (2011)
[143] Nordhausen, K., Ollila, E., Oja, H.: On the performance indices of ICA and blind source separation. In: Proceedings of IEEE 12th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2011), pp. 486–490 (2011)
[144] Salokangas, R.K., Helminen, M., Koivisto, A.M., Rantanen, H., Oja, H., Pirkola, S., Wahlbeck, K., Joukamaa, M.: Incidence of hospitalised schizophrenia in Finland since 1980: decreasing and increasing again. Soc. Psychiatry Psychiatr. Epidemiol. **46**(4), 343–350 (2011)
[145] Tokola, K., Larocque, D., Nevalainen, J., Oja, H.: Power, sample size and sampling costs for clustered data. Stat. Probab. Lett. **81**(7), 852–860 (2011)
[146] Datta, S., Nevalainen, J., Oja, H.: A general class of signed-rank tests for clustered data when the cluster size is potentially informative. J. Nonparametr. Stat. **24**, 797–808 (2012)
[147] Ilmonen, P., Oja, H., Serfling, R.: On invariant coordinate system (ICS) functionals. Int. Stat. Rev. **80**(1), 93–110 (2012)

[148] Lemponen, R., Larocque, D., Nevalainen, J., Oja, H.: Weighted rank tests and Hodges-Lehmann estimates for the multivariate two-sample location problem with clustered data. J. Nonparametr. Stat. **24**, 977–991 (2012)
[149] Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S.: Statistical properties of a blind source separation estimator for stationary time series. Stat. Probab. Lett. **82**, 1865–1873 (2012)
[150] Nordhausen, K., Gutch, H.W., Oja, H., Theis, F.J.: Joint diagonalization of several scatter matrices for ICA. In: Theis, F.J., Cichocki, A., Yeredor, A., Zibulevsky, M. (eds.) Latent Variable Analysis and Signal Separation - 10th International Conference, LVA/ICA 2012, Tel Aviv, Israel, 12–15 March 2012. Proceedings of Lecture Notes in Computer Science, vol. 7191, pp. 172–179. Springer (2012)
[151] Oja, H.: Descriptive statistics for nonparametric models. The impact of some Erich Lehmann's papers. In: Rojo, J. (ed.) Selected Works of E. L. Lehmann, pp. 451–457. Springer, New York (2012)
[152] Oja, H., Nordhausen, K.: Independent component analysis. In: El-Shaarawi, A.-H., Piegorsch, W. (eds.) Encyclopedia of Environmetrics, 2nd edn., pp. 1352–1360. Wiley, Chichester (2012)
[153] Taskinen, S., Koch, I., Oja, H.: Robustifying principal component analysis with spatial sign vectors. Stat. Probab. Lett. **82**, 765–774 (2012)
[154] Kiwuwa-Muyingo, S., Oja, H., Walker, A.S., Ilmonen, P., Levin, J., Mambule, I., Mugyenyi, P., Todd, J., DART Trial team. Dynamic logistic regression model and population attributable fraction to investigate the association between adherence, missed visits and mortality: a study of HIV-infected adults surviving the first year of ART. BMC Infect. Dis. **13**(395) (2013)
[155] Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S.: Fast equivariant JADE. In: Proceedings of 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013), pp. 6153–6157 (2013)
[156] Oja, H.: Multivariate median. In: Becker, C., Fried, R., Kuhnt, S. (eds.) Robustness and Complex Data Structures, pp. 3–15. Springer, Berlin (2013)
[157] Fischer, D., Oja, H., Schleutker, J., Sen, P.K., Wahlfors, T.: Generalized Mann-Whitney type tests for microarray experiments. Scand. J. Stat. **41**, 672–692 (2014)
[158] Liski, E., Nordhausen, K., Oja, H.: Supervised invariant coordinate selection. Stat. J. Theor. Appl. Stat. **48**, 711–731 (2014)
[159] Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S.: Deflation-based separation of uncorrelated stationary time series. J. Multivar. Anal. **123**, 214–227 (2014)
[160] Nevalainen, J., Datta, S., Oja, H.: Inference on the marginal distribution of clustered data with informative cluster size. Stat. Pap. **55**, 71–92 (2014)
[161] Tokola, K., Lundell, A., Nevalainen, J., Oja, H.: Design and cost optimization for hierarchical data. Statistica Neerlandica **68**, 130–148 (2014)

A.2 R Packages of Hannu Oja

[R1] Nordhausen, K., Oja, H., Tyler, D.E.: ICS: Tools for Exploring Multivariate Data via ICS/ICA (2007) http://www.cran.r-project.org/package=ICS
[R2] Nordhausen, K., Sirkiä, S., Oja, H., Tyler, D.E.: ICSNP: Tools for Multivariate Nonparametrics (2007) http://www.cran.r-project.org/package=ICSNP
[R3] Nordhausen, K., Cardoso, J.F., Miettinen, J., Oja, H., Ollila, E., Taskinen, S.: JADE: JADE and other BSS methods as well as some BSS performance criteria (2007) http://www.cran.r-project.org/package=JADE
[R4] Sirkiä, S., Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S.: SpatialNP: Multivariate nonparametric methods based on spatial signs and rank (2009) http://www.cran.r-project.org/package=SpartialNP

[R5] Nordhausen, K., Möttönen, J., Oja, H.: MNM: Multivariate Nonparametric Methods. An Approach Based on Spatial Signs and Rank (2009) http://www.cran.r-project.org/package=MNM
[R6] Fischer, D., Oja, H.: gMWT: Generalized Mann-Whitney type tests based on probabilistic indices and new diagnostic plots (2012) http://www.cran.r-project.org/package=gMWT
[R7] Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S.: fICA: Classic, reloaded and adaptive FastICA algorithms (2013) http://www.cran.r-project.org/package=fICA
[R8] Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S.: BSSasymp: Asymptotic covariance matrices of some BSS mixing and unmixing matrix estimates (2013) http://www.cran.r-project.org/package=BSSasymp
[R9] Liski, E., Nordhausen, K., Oja, H., Ruiz-Gazen, A.: LDRTools: Tools for Linear Dimension Reduction (2014) http://www.cran.r-project.org/package=LDRTools

A.3

Table 2.2 All authors in network G, their cluster memberships and labels for plotting

Label	Author	CL	Label	Author	CL	Label	Author	CL
5	Brown	1	91	Moring	2	34	Ilonen	6
6	Busarova	1	94	Muhli	2	48	Kenward	6
7	Chakraborty	1	96	Naarala	2	55	Knip	6
8	Chaudhuri	1	108	Pirilä	2	64	Korhonen	6
9	Critchley	1	111	Pykäläinen	2	100	Niinistö	6
12	Datta	1	122	Seitamaa	2	125	Simell	6
13	Dümbgen	1	129	Sorri	2	130	Sundvall	6
17	Ghosh	1	136	Tarvainen	2	145	Uusitalo	6
18	Gutch	1	139	Tienari	2	148	Veijola	6
20	Hallin	1	155	Wahlberg	2	151	Virtanen	6
22	Hartikainen	1	159	Wynne	2	32	Ilmonen	7
26	Hettmansperger	1	4	Bracken	3	53	Kiwuwa-Muyingo	7
29	Hüsler	1	16	Fischer	3	77	Levin	7
61	Koivunen	1	31	Ikonen	3	85	Mambule	7
65	Koshevoy	1	37	Isotalo	3	86	Mandal	7
67	Krause	1	38	Jalava	3	93	Mugyenyi	7
69	Läärä	1	74	Lehmusvaara	3	124	Serfling	7
73	Larocque	1	87	Mattila	3	140	Todd	7
75	Lemponen	1	115	Rauhala	3	157	Walker	7
79	Liski	1	120	Schindler	3	25	Helminen	8
80	Lundell	1	121	Schleutker	3	44	Joukamaa	8
81	Luukkonen	1	123	Sen	3	59	Koivisto	8
92	Möttönen	1	135	Tammela	3	109	Pirkola	8
97	Nadar	1	149	Vihinen	3	114	Rantanen	8
98	Nevalainen	1	152	Visakorpi	3	118	Salokangas	8
99	Niinimaa	1	156	Wahlfors	3	154	Wahlbeck	8
101	Nordhausen	1	11	Croux	4	35	Ilveskoski	9

(continued)

Table 2.2 (continued)

Label	Author	CL	Label	Author	CL	Label	Author	CL
102	Nyblom	1	15	Eriksson	4	45	Kähonen	9
103	Oja	1	27	Hillary	4	63	Kööbi	9
104	Ollila	1	40	J. Miettinen	4	66	Koskela	9
105	Orponen	1	46	Kankainen	4	76	Leskinen	9
106	Paindaveine	1	56	Koch	4	95	Mustonen	9
110	Pörsti	1	88	McCann	4	133	Tahvanainen	9
116	Ronkainen	1	89	McGuinness	4	78	Lindroos	10
132	Tableman	1	107	Paton	4	82	Mäenpää	10
134	Talvisara	1	112	Randles	4	128	S. Miettinen	10
138	Theis	1	119	Sandberg	4	131	Suuronen	10
141	Tokola	1	127	Sirkiä	4	160	Ylikomi	10
142	Topchii	1	137	Taskinen	4	24	Hasan	11
144	Tyurin	1	143	Tyler	4	28	Himanen	11
153	Visuri	1	10	Croudace	5	30	Huupponen	11
2	Alho	2	14	ElSaid	5	117	Saastamoinen	11
3	Anais-Tanner	2	23	Hartikainen-Sorri	5	147	Värri	11
19	Hakko	2	36	Isohanni	5	47	Keinänen-Kiukaanniemi	12
33	Ilo	2	39	Järvelin	5	51	Kiultu	12
42	Jokinen	2	41	Jokelainen	5	52	Kiuttu	12
49	Keskitalo	2	43	Jones	5	72	Larivaara	12
50	Kilkku	2	57	Koiranen	5	146	Väisänen	12
58	Koistinen	2	83	Mäkikyrö	5	21	Härkänen	13
60	Koivu	2	90	Moilanen	5	54	Knekt	13
62	Kolassa	2	113	Rantakallio	5	68	Laaksonen	13
70	Lahti	2	126	Sipilä	5	150	Virtala	13
71	Läksy	2	158	Wendt	5			
84	Mäki-Torkko	2	1	Alfthan	6			

The authors are sorted according to the cluster membership and alphabetically within the clusters

References

Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal Complex Syst. **1695** (2006)

Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)

Francois, R.: bibtex: bibtex parser. R package version 0.4.0. http://CRAN.R-project.org/package=bibtex (2014)

Kolaczyk, E.D., Csardi, G.: Statistical Analysis of Network Data with R. Springer, New York (2014)

Kretschmer, H.: Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility in the Web. Scientometrics **60**(3), 409–420 (2004)

Lancichinetti, A., Fortunato, S.: Community detection algorithms: a comparative analysis. Phys. Rev. E **80**, 056117 (2009)

Mardsen, P.V.: Egocentric and sociocentric measures of network centrality. Soc. Networks **24**, 407–422 (2002)

Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E **74**, 036104 (2006)

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2014)

South, A.: rworldmap: a new R package for mapping global data. R J. **3**(1), 35–43 (2011)

Part II
Univariate Nonparametric and Robust Methods

Chapter 3
Approximate U-Statistics for State Waiting Times Under Right Censoring

Somnath Datta, Douglas J. Lorenz, and Susmita Datta

Abstract We develop two different adaptations of a U-statistic based on the waiting times in a given transient state in a multistate system when the state entry and/or the exit times are subject to right censoring. In the first version, the inverse probability of censoring weights calculated based on the state exit times are used along with m-tuples of fully observed waiting times, m being the degree of the kernel of the U-statistic. In the second version, an approximate statistic is defined as a multiple integral with respect to a product of Satten–Datta estimators of a state waiting time survival function. We provide a simulation study to investigate the finite sample behavior of the statistics. We demonstrate that the second version is more efficient since it utilizes additional data where the exit times may be censored. The asymptotic normality of our estimators is also studied through simulation. We further extend our approximate U-statistics to that of a K-sample U-statistic of waiting times under right censoring. Another extension considers waiting time data that are clustered. We apply our U-statistics to test whether initial functional status has a significant impact on the waiting time of an intermediate state of functional recovery of a spinal cord injured patient.

Keywords Dependent censoring • Multistate models • Sojourn times • Survival analysis

3.1 Introduction

U-statistics, as introduced by Hoeffding (1948), are useful tools in constructing various estimators and test statistics, including nonparametric tests (Serfling 1980). The asymptotics for U-statistics are generally studied via the Hoeffding decomposition

S. Datta (✉) • S. Datta
Department of Biostatistics, University of Florida, Gainesville, FL, USA
e-mail: somnath.datta@ufl.edu; susmita.datta@ufl.edu

D.J. Lorenz
Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, USA
e-mail: douglas.lorenz@louisville.edu

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_3

(Bickel and Lehmann 1979; Randles and Wolfe 1979; Serfling 1980; Sen 1981; Lee 1990).

For complete (e.g., uncensored or non-missing) data, U-statistics are obtained by averaging a kernel computed at various sub-samples of the original sample, thereby extending the notion of a sample mean. The kernel h of a U-statistic is generally assumed to be a symmetric function of its m arguments ($1 \leq m < n$), known as its order or degree. Thus, if $X_1, \cdots, X_n$ denote the complete data sample, the corresponding U -statistic is given by

$$U = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_m \leq n} h(X_{i_1}, \ldots, X_{i_m}). \tag{3.1}$$

Assuming that the X_i are independent and identically distributed (i.i.d.) with a common marginal distribution function F, the corresponding functional for which U is unbiased and asymptotically consistent is given by

$$\theta_F = Eh(X_1, \ldots, X_m) = \int_{\mathscr{X}^m} h(x_1, \cdots, x_m) dF(x_1) \cdots dF(x_m). \tag{3.2}$$

Extensions of U -statistics to various incomplete data problem have received attention in recent years (Akritas 1986; Gijbels and Veraverbeke 1991; Stute and Wang 1993a,b; Stute 1995; Bose and Sen 1999, 2002; Schisterman and Rotnizky 2001; Datta et al. 2010; Fan and Datta 2013). Such extensions are often mathematically challenging; however, more importantly they are useful from a practical standpoint since in various real-life situations, notably in biomedical applications, data are often incomplete due to missingness and/or censoring. Thus, these extensions provide a user with a familiar and powerful methodology for handling various inferential issues with such data. Typically, such extended forms of U-statistics are not technically U-statistics in the classical sense. However, generally speaking they are asymptotically equivalent to a full data U-statistic at least in the sense of approximate unbiasedness for a corresponding population parameter θ. Their variances are generally more complex. Rather than finding exact linearization and closed-form estimates of the asymptotic variances one may attempt techniques such as the jackknife or bootstrap (see, e.g., Efron 1982) to compute the variance and perform large sample inference for θ using these extended U-statistics.

The rest of this chapter is organized as follows. In Sect. 3.2, we introduce our censored U-statistic for waiting times under independent right censoring of the event times so that both state entry and exit times are potentially right censored. We also extend its definition to the case of dependent right censoring when there are observable covariates that explain the dependence between the true event times and the censoring times. Section 3.3 presents simulation results comparing the two versions of the extended U-statistics, and demonstrates the effectiveness of jackknife variance estimates and approximate normality of the sampling distribution of our U-statistics. Useful extensions of our U-statistics are presented in Sect. 3.4, including extensions to the multi-sample case and also to clustered multistate data with potentially informative cluster size. A numerical illustration is provided in

Sect. 3.5, where we compare certain state waiting times of spinal cord injury (SCI) patients. The chapter ends with a brief discussion in Sect. 3.6.

3.2 IPCW U-statistics for Right Censored Waiting Times

We consider n individuals in a multistate system. We concentrate on an intermediate (e.g., transient state) such that all individuals would eventually pass through that state. For the ith individual, let X_i^*, V_i^*, and $W_i^* = V_i^* - X_i^*$ be the state entry, exit, and waiting times, respectively. In the presence of right censoring by C_i, each of these times may be unobserved. Of course, if V_i^* is observed (i.e., when $V_i^* \leq C_i$), the rest are observed as well. However, V_i^* and W_i^* may be unavailable even if X_i^* is observed, which happens if $X_i^* \leq C_i < V_i^*$. One has to handle each case of censoring appropriately to avoid selection bias while maintaining efficiency.

Our observed data consist of the four-tuples $(X_i, \xi_i, V_i, \delta_i), 1 \leq i \leq n$, where $X_i = \min(X_i^*, C_i)$ and $V_i = \min(V_i^*, C_i)$ are the (right) censored state entry and exit times, and $\xi_i = I(X_i^* \leq C_i)$ and $\delta_i = I(V_i^* \leq C_i)$ are censoring indicators for stage entry and exit for the ith subject. Also, let $\overline{\xi}_i = 1 - \xi_i$ and $\overline{\delta}_i = 1 - \delta_i$. Define the censored waiting times as $W_i = V_i - X_i$. Note that W_i is computable from the observed data and equals W_i^* if and only if $\delta_i = 1$. Let F_{W^*} be the waiting time distribution function and $S_{W^*} = 1 - F_{W^*}$ its survival function.

We define two different versions of our censored data U-statistics for waiting times. The first is based on data values observed up to at least time V_i^*, i.e., based on the waiting times of individuals with $\delta_i = 1$. The second approach we propose is to express the U-statistic as an integral with respect to the product of estimated state waiting time distribution functions $\hat{F}_{W^*}$. The estimated state waiting time distribution function was proposed by Satten and Datta (2002) and is reviewed in the next subsection for completeness. In the first approach, the selection bias of using δ_i is removed by reweighting the term by the inverse of its conditional expectation given the observed data. These and other weights were also used in constructing the Satten–Datta estimator $\hat{F}_{W^*}$. Computation of these weights depends on the assumed model of the censoring hazard, the simplest of which is that of random censoring. Details are covered in the next subsection.

3.2.1 Random Right Censoring

Suppose (X_i^*, V_i^*, C_i), $1 \leq i \leq n$, are i.i.d. and that the censoring mechanism is independent of the multistate process, and thus for each i, C_i is independent of $\{X_i^*, V_i^*\}$. Note that, in this case $E(\delta_i | X_i^*, V_i^*) = E(\delta_i | V_i^*) = S_C(V_i^* -)$, where $-$ denotes left limit and S_C is the common survival function of the censoring random variables. Thus, extending the mean-preserving reweighting approach of Koul et al.

(1981) or Datta (2005) for the sample mean, we could define our censored waiting time U-statistic based on a kernel h as

$$U_c = \frac{1}{\binom{n}{m}} \sum_i \frac{h(W_{i_1}, \cdots, W_{i_m}) \prod_{\ell \in i} \delta_\ell}{\prod_{\ell \in i} S_C(V_i-)}, \tag{3.3}$$

where the notation $\ell \in i$ is used to indicate that ℓ is one of the integers $\{i_1, \cdots, i_m\}$. Note that U_c itself is a U-statistic based on the triplets (W_i, V_i, δ_i), $1 \le i \le n$ with kernel H

$$H\Big((W_1, V_1, \delta_1), \ldots, (W_m, V_m, \delta_m)\Big) = \frac{h(W_1, \cdots, W_m) \prod_{\ell=1}^m \delta_\ell}{\prod_{\ell=1}^m S_C(V_i-)}. \tag{3.4}$$

Furthermore,

$$\begin{aligned}
E\left(H\Big((W_1, V_1, \delta_1), \ldots, (W_m, V_m, \delta_m)\Big)\right) &= \\
&= E\left\{\frac{h(W_1^*, \cdots, W_m^*) \prod_{\ell=1}^m \delta_\ell}{\prod_{\ell=1}^m S_C(V_i^*-)}\right\} \\
&= E\left[\frac{h(W_1^*, \cdots, W_m^*)}{\prod_{\ell=1}^m S_C(V_i^*-)} E\left\{\prod_{\ell=1}^m \delta_\ell \,\middle|\, X_1^*, V_1^*, \cdots, X_m^*, V_m^*\right\}\right] \\
&= E\left[\frac{h(W_1^*, \cdots, W_m^*)}{\prod_{\ell=1}^m S_C(V_i^*-)} \prod_{\ell=1}^m E\{\delta_\ell | X_\ell^*, V_\ell\}\right], \\
&= E\Big(h(W_{i_1}^*, \cdots, W_{i_m}^*)\Big) = \theta,
\end{aligned}$$

provided $S_c(V_i^*-) > 0$, for each i, with probability 1.

Unfortunately, U_c as defined above is not usable, as S_c is unknown and thus not a legitimate statistic. We can estimate S_c by the Kaplan–Meier estimator, where the role of censored and uncensored exit times is reversed. Replacing S_c by its estimated version, we get our first approximate U-statistic for right censored waiting times,

$$\hat{U}(1) = \frac{1}{\binom{n}{m}} \sum_i \frac{h(W_{i_1}, \cdots, W_{i_m}) \prod_{\ell \in i} \delta_\ell}{\prod_{\ell \in i} \hat{S}_C(V_i-)}. \tag{3.5}$$

We refer to this statistic an inverse probability of censoring weighted (IPCW) U-statistic based on censored waiting times although technically, it is not a U-statistic since $\hat{S}_c$ uses data from all individuals. Nevertheless, it serves the same purpose as the corresponding U-statistic for the full data and is consistent for the same $\theta(F)$. One can obtain its asymptotic distribution following a censored data U-statistic result of Datta et al. (2010) where one identifies the marks with the waiting times W^* and the failure times T^* with the state exit time V^*. While this statistic is consistent and asymptotically normal, its asymptotic variance is larger than that of a U -statistic

based on full data waiting times (W^*). This efficiency loss due to censoring is to be expected.

Next, we introduce another censored waiting time U-statistic that makes more efficient use of the available data, as some of the observations with the observed entry times but censored exit times are used in its computation. This situation does not arise with right censored failure times and thus this U-statistic was not covered in Datta et al. (2010). Since the population quantity being estimated takes integral form $\theta = \int h(x_1, \cdots, x_m) dF_W(x_1) \cdots dF_W(x_m)$ we can define an approximate U-statistic by replacing F_W by a nonparametric estimator. Satten and Datta (2002) proposed a Kaplan–Meier type nonparametric estimator of F_W which we now review.

Define the following reweighted versions of the "counting" and "number at risk" processes corresponding to the right censored weighting times:

$$N_c^W(w) = \sum_{i=1}^{n} \frac{\delta_i I(W_i \le w)}{S_c(V_i-)} \tag{3.6}$$

and

$$Y_c^W(w) = \sum_{i=1}^{n} \frac{\xi_i I(W_i \ge w)}{S_c(X_i + w-)}. \tag{3.7}$$

It is straightforward to show, using previously suggested arguments, that $EN_c^W(w) = EN^{W^*}(w)$, where $N^{W^*}(w) = \sum_{i=1}^{n} I(W_i^* \le w)$ counts the number of full data waiting times not exceeding w. Note that on the set $\xi_i = 1$, the entry time is observed and hence $X_i = X_i^*$. Moreover, $\xi_i I(W_i \ge w) = I(W_i^* \ge w) I(C_i \ge X_i^* + w-)$, and hence $EY_c^W(w) = EY^{W^*}(w)$, where $Y^{W^*}(w) = \sum_{i=1}^{n} I(W_i^* \ge w)$ is the "number at risk" process corresponding to the full data waiting times. The Satten–Datta estimator is obtained by a Kaplan–Meier formula where estimated versions of these two reweighted processes are used:

$$\hat{F}_W(w) = 1 - \prod_{s \le w} \left(1 - \frac{\Delta \hat{N}_c^W(s)}{\hat{Y}_c^W(s)}\right), \tag{3.8}$$

where $\hat{N}_c^W$ and $\hat{Y}_c^W$ are obtained by replacing S_c by its Kaplan–Meier estimator $\hat{S}_c$ in (6) and (7), respectively; Δ denotes the jump of a process.

Finally, our second approximate U-statistic is given by

$$\hat{U}(2) = \int h(w_1, \cdots, w_m) d\hat{F}_W(w_1) \cdots d\hat{F}_W(w_m). \tag{3.9}$$

Note that for complete data, such integrals with respect to the empirical distribution are referred to as V-statistics, which are asymptotically equivalent to the corresponding U-statistics. Since, in presence of censoring, we could only define approximate U-statistics, we are not making such a fine distinction. However if desired, one can

obtain another version by expanding (9) into a sum and only retaining the terms over distinct indices.

Since the version $\hat{U}(2)$ uses the Datta–Satten estimator which in turn uses data points with $\xi = 1$, rather than $\delta = 1$, in computing the "number at risk" process, we expect it to be more efficient than $\hat{U}(1)$. In the simulation section we compare the behavior of $\hat{U}(1)$ to that of $\hat{U}(2)$ in several settings. We show that while both $\hat{U}(1)$ and $\hat{U}(2)$ seem to be nearly unbiased, $\hat{U}(2)$ has lower variance than $\hat{U}(1)$.

In the following subsection, we show how to modify these estimators for the case of dependent censoring, where the censoring hazard is modeled through a set of observable subject-specific baseline and/or time-varying covariates that explain the dependence between censoring and the multistate process.

3.2.2 Dependent Censoring

The assumption of random (or independent) censoring can be relaxed by modeling the censoring process via auxiliary information. The resulting U-statistics broaden the scope of their applicability. This approach of modeling dependent censoring through a collection of observable covariables (fixed or time varying) $\mathbf{Z} = \{\mathbf{Z}(t) : t \geq 0\}$ has been advocated in Robins and Rotnizky (1992), Robins (1993), as well as by us in a number of prior publications (Datta and Satten 2002; Satten et al. 2001b; Satten and Datta 2002, 2004). Satten and Datta (2001) showed that the standard Kaplan–Meier estimator under independent censoring is also a special case of IPCW. In what follows, we assume that the covariates $\mathbf{Z}$ are observable for all individuals (including those whose exit times are censored) at least up to the time V. The following technical condition is equivalent to that of "no unmeasured confounders" and means that the covariables $\mathbf{Z}$ explain all the dependence between the entry/exit times and censoring times:

$$\lim_{dt \to 0} \frac{P\{t \leq C < t + dt | \mathbf{Z}(u), 0 \leq u < t, V \geq t, X^*, V^*\}}{dt} = \\ = \lim_{dt \to 0} \frac{P\{t \leq C < t + dt | \mathbf{Z}(u), 0 \leq u < t, V \geq t\}}{dt} \doteq \lambda_c(t). \tag{3.10}$$

Nonparametric estimation of the integrated censoring hazard $\Lambda_c(t) = \int_0^t \lambda_c(s)ds$ is discussed below. Once an estimator of $\Lambda_c(t)$ is obtained we can redefine $\hat{S}_c(t)$ as $\hat{S}_c(t) = \prod_{s \leq t}(1 - \hat{\Lambda}_c(s)ds)$ or $\hat{S}_c(t) = \exp\{-\hat{\Lambda}_c(t)\}$. Note that in general, $\hat{S}_c(t)$ will depend on the subject-specific covariates $\mathbf{Z}_i$; in other words, we have suppressed the subject index i for notational simplicity. Finally, the modified approximate U-statistics $\hat{U}(1)$ and $\hat{U}(2)$ are obtained by the same formulas as in the previous section, where we use the current definition of $\hat{S}_c$. These modified U-statistics will provide asymptotically unbiased estimators of $\theta(F_W)$ even in the presence of dependent censoring.

Suppose, we have a single time independent (i.e., baseline) covariate Z taking finitely many values $\{z_1, \cdots, z_K\}$ that explains the dependence between C and $\{X^*, V^*\}$. In this case, we can split the sample into K groups by the value taken by Z. That is, let $\mathscr{S}_k = \{i : Z_i = z_k, 1 \le i \le n\}, 1 \le k \le K$. Let $\hat{S}_{c,k}$ be the Kaplan–Meier estimator of censoring times based on $\mathscr{S}_k$, i.e.,

$$\hat{S}_{c,k}(t) = \prod_{s \le t} \left(1 - \frac{\Delta N_{c,k}(s)}{Y_{c,k}(s)}\right), \ 1 \le k \le K,$$

where $N_{c,k}(s) = \sum_{i \in \mathscr{S}_k} I(C_i \le s, \delta_i = 0)$ and $Y_{c,k}(s) = \sum_{i \in \mathscr{S}_k} I(V_i \ge s)$. In this case, $\hat{S}_c = \hat{S}_{c,k}$, if $Z_i = z_k$.

Next consider the situation when $Z(t)$ is the state occupied just prior to time t, and we would be assuming that the censoring hazard may be different in different states of a multistate system with K states. Using a common Kaplan–Meier estimator for the overall censoring distribution will lead to biased results in this case. Thus we compute K separate Kaplan–Meier estimators corresponding to K different hazards as before, where $\Delta N_{c,k}(s) = \sum_{i=1}^n I(C_i = s, \delta_i = 0, Z_i(s) = k)$ and $Y_{c,k}(s) = \sum_{i=1}^n I(V_i \ge s, Z_i(s) = k)$. For a given individual, let $\{\tau_j, j \ge 1\}$ be the collection of its transition times, and let $s(t) = i_k$, for $\tau_{k-1} \le t < \tau_k, k \ge 1, \tau_0 = 0$. Then $\hat{S}_c$ is given by

$$\hat{S}_c(t) = \left\{\prod_{l=1}^{k-1} \frac{\hat{S}_{c,i_l}(\tau_l)}{\hat{S}_{c,i_l}(\tau_{l-1})}\right\} \frac{\hat{S}_{c,i_k}(t)}{\hat{S}_{c,i_k}(\tau_{k-1})}, \text{ for } \tau_{k-1} \le t < \tau_k. \tag{3.11}$$

The above two examples are special cases of Aalen's linear hazard model (Aalen 1980, 1989). This is a nonparametric model yielding a flexible choice for computation of the weights in a variety of situations. Mathematically, the censoring hazard is modeled as

$$\lambda_c(t) = \sum_{k=0}^{p} \beta_k(t) W_k(t), \tag{3.12}$$

where the $W_k(t) = \phi_k(\mathbf{Z}(u), 0 \le u < t)$, are predictable functions of the time dependent covariates $\mathbf{Z}(t)$, each $\beta_k(t)$ is an unknown function, and where one assumes that the first component $W_0(t) \equiv 1$. In the two examples mentioned above, $p = K$, the β_ks denote the K hazards ($\beta_0 \equiv 0$), $W_k(t) = I(Z = z_k)$ in the first example, and $W_k(t) = I(s(t-) = k)$ in the second example, where $s(t-)$ is the state occupied at time $t-$. Letting $B_k(t) = \int_0^t \beta_k(s) ds$, Aalen's estimator of the vector $B(t) = \big(B_0(t), \cdots, B_p(t)\big)^\top$ is given by

$$\widehat{\mathbf{B}}(t) = \sum_{i=1}^{n} I(V_i \le t)(1 - \delta_i) \mathbf{A}^{-1}(t) \mathbf{W}_i(t) \tag{3.13}$$

where $\mathbf{W}_i(t) = \left(W_{i0}(t), W_{i1}(t), \cdots, W_{ip}(t)\right)^\top$ and the matrix $\mathbf{A}(t)$ is given by

$$\mathbf{A}(t) = \sum_{i=1}^{n} I(V_i \geq t)\mathbf{W}_i(t)\mathbf{W}_i^\top(t). \tag{3.14}$$

This leads to the subject specific estimates of cumulative hazards

$$\begin{aligned}\hat{\Lambda}_c^i(t) &= \sum_{k=0}^{p} \int_0^t W_{ik}(t) d\hat{B}_k(t) \\ &= \sum_{j=1}^{n} I(V_j \leq t)(1-\delta_j)\mathbf{W}_i^\top(V_j)\widehat{\mathbf{A}}^{-1}(V_j)\mathbf{W}_j(V_j)\,, \quad t \leq V_i. \end{aligned} \tag{3.15}$$

Finally, we can use $\hat{S}_{c,i}(t) = exp\{-\hat{\Lambda}_c^i(t)\}$ or $\hat{S}_{c,i}(t) = \prod_{s\leq t}\{1-d\hat{\Lambda}_c^i(s)\}$ in defining our reweighted estimators.

3.3 Simulation Study

To evaluate the estimators $\hat{U}(1)$ and $\hat{U}(2)$, we conducted a simulation study on a irreversible three-state model. The natural logarithm of the entry (X_i^*) and waiting times (W_i^*) in the transient model state for $n = 500$ individuals were generated from the bivariate normal distribution with marginal means equal to 0, marginal variances equal to 1, and covariance equal to ρ, where ρ varied between -0.5 and 0.5 for different simulation settings. Censoring times were generated from the Weibull distribution with shape parameter 2 and scale parameters 2 and 5 to represent heavy and light censoring, respectively. Under these parameter settings, approximately 64 % and 87 % of individuals enter the transient model state before censoring ($\xi = 1$) and 36 % and 69 % of individuals exit before censoring ($\delta = 1$) under heavy and light censoring, respectively.

For each of 5,000 Monte Carlo iterations, we calculated $\hat{U}(1)$ and $\hat{U}(2)$ based on the order-2 kernel $h(w_1, w_2) = I(\log(w_1) + \log(w_2) > 0)$, for which we note that $\theta = E(h(w_1, w_2)) = 0.5$. In calculating $\hat{U}(1)$ and $\hat{U}(2)$, we used the Kaplan–Meier estimator of the censoring survival function in formulas (3.5), (3.6), and (3.7). We calculated the bias of $\hat{U}(1)$ and $\hat{U}(2)$ as the difference between the empirical average of the 5,000 replicate estimates and the true value of 0.5. We calculated jackknife estimates of the variance of $\hat{U}(1)$ and $\hat{U}(2)$ and used these to calculate coverage probabilities for normal asymptotic confidence intervals of varying size.

We noted that both $\hat{U}(1)$ and $\hat{U}(2)$ were approximately unbiased and that the jackknife variance of $\hat{U}(2)$ tended to be lower than that of $\hat{U}(1)$ (Table 3.1). Both $\hat{U}(1)$ and $\hat{U}(2)$ exhibited confidence interval coverage in reasonable correspondence with nominal levels (Fig. 3.1). $\hat{U}(1)$ tended to be slightly conservative at lower

Table 3.1 Average empirical bias, jackknife variance, and 95 % confidence interval coverage probabilities for $\hat{U}(1)$ and $\hat{U}(2)$ for various simulation settings

Censoring	Correlation	Statistic	Bias	Variance	Coverage probability
Light	0.5	$\hat{U}(1)$	0.0009	0.0018	0.9506
		$\hat{U}(2)$	−0.0005	0.0014	0.9496
Light	−0.5	$\hat{U}(1)$	0.0011	0.0017	0.9552
		$\hat{U}(2)$	−0.0012	0.0016	0.9526
Heavy	0.5	$\hat{U}(1)$	−0.0007	0.0028	0.9314
		$\hat{U}(2)$	0.0010	0.0020	0.9498
Heavy	−0.5	$\hat{U}(1)$	0.0009	0.0022	0.9344
		$\hat{U}(2)$	−0.0008	0.0016	0.9478

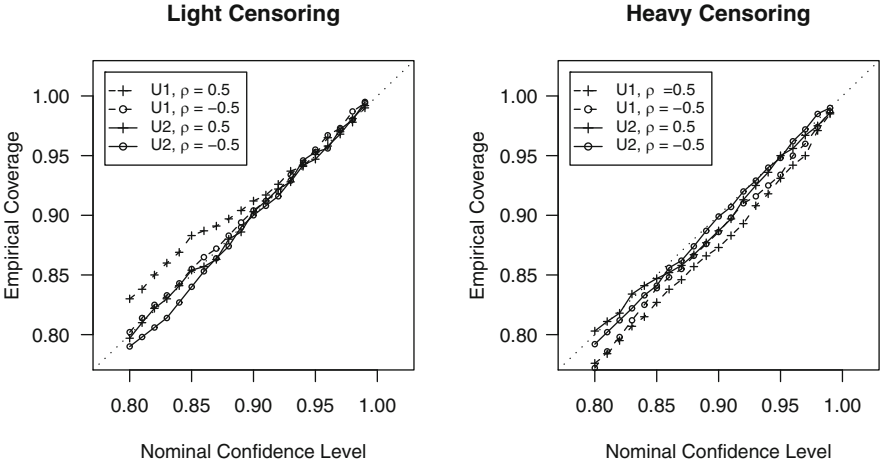

Fig. 3.1 Plot of empirical confidence levels for $\hat{U}(1)$ and $\hat{U}(2)$ against nominal levels for each simulation setting

confidence levels ($< 90\,\%$) when censoring was light and state waiting times were positively correlated. $\hat{U}(1)$ also exhibited slight undercoverage under heavy censoring, particularly at lower confidence levels ($< 90\,\%$), whereas $\hat{U}(2)$ exhibited coverage closer to nominal levels under heavy censoring. This is a likely product of the use $\hat{U}(2)$ makes of observations that are censored in the transient model state when approximating the number at risk of exit from the transient model state, i.e., those with $\xi = 1$ but $\delta = 0$.

3.4 Extensions

In this section, we introduce two useful extensions of our waiting time U-statistics.

3.4.1 K-sample Approximate U-statistics for Right Censored Waiting Times

Consider K independent samples of waiting times from K multistate processes $\{W^*_{k,i}, 1 \leq k \leq n_i\}, 1 \leq i \leq K$. A K-sample (extended) U-statistic based on a kernel h of order $(m_1, \cdots, m_K)$ is defined as

$$U = \frac{1}{\prod_{k=1}^{K} \binom{n_k}{m_k}} \sum_{i(\cdot,\cdot)} h(W^*_{1,i(1,1)}, \cdots, W^*_{1,i(m_1,1)}; \cdots; W^*_{K,i(K,1)}, \cdots, W^*_{1,i(K,m_K)}), \tag{3.16}$$

where $1 \leq i(k,1) < \cdots < i(k,m_k) \leq n_k,\ 1 \leq k \leq K$.

Suppose we have right censored entry and exit time samples from K independent multistate processes $\{X_{k,i}, V_{k,i}, \delta_{k,i}, \xi_{k,i}, Z_{k,i} : 1 \leq i \leq n_i\}, 1 \leq k \leq K$. Then the two versions of our approximate K sample U-statistics will be given by

$$\hat{U}(1) = \frac{1}{\prod_{k=1}^{K} \binom{n_k}{m_k}} \sum_{i(\cdot,\cdot)} \frac{\prod_{k=1}^{K} \prod_{j=1}^{m_k} \delta_{k,i(k,j)}}{\prod_{k=1}^{K} \prod_{j=1}^{m_k} \hat{S}_c(V_{k,i(k,j)}-)} \times h(W_{1,i(1,1)}, \cdots, W_{1,i(m_1,1)}; \cdots; W_{K,i(K,1)}, \cdots, W_{1,i(K,m_K)}) \tag{3.17}$$

and

$$\hat{U}(2) = \int h(w_{1,i(1,1)}, \cdots, w_{1,i(1,m_1)}; \cdots; w_{K,i(K,1)}, \cdots, w_{1,i(K,m_K)}) dF_1\left(w_{1,i(1,1)}\right) \cdots dF_1\left(w_{1,i(m_1,1)}\right) \cdots dF_K\left(w_{K,i(K,1)}\right) \cdots dF_K\left(w_{1,i(K,m_K)}\right), \tag{3.18}$$

respectively. Once again, jackknife can be used to calculate the variance of a $k-$sample statistic; see, e.g., Arvesen (1969) or Schechtman and Wang (2004).

3.4.2 Approximate Marginal U-statistics for Clustered Waiting Times

In many applications, data come in the form of independent clusters such that observations within a cluster are potentially correlated. We assume that we have a collection of waiting times $\{W^*_{ij} : 1 \leq j \leq N_i, 1 \leq i \leq M\}$, where i denotes clusters and j denotes an observation within a cluster. Thus M denotes the total number of clusters and N_i denotes the number of observations in cluster i. For example, Lorenz and Datta (2015) considered state waiting time data of spinal cord injured patients in an activity-based rehabilitation program who go through various recovery stages before they are discharged. One may be interested in computing a one sample U-

statistic corresponding to a marginal distribution of the collection of state waiting times of a patient or compare the marginal distributions of the collection of state waiting times for patients belonging to two different groups, say, based on their injury characteristics. The waiting times at various stages of functional recovery for the same patients are likely to be correlated. Thus, for this example, an individual patient represents a single cluster.

Often, inference for marginal distributions of clustered data is of interest. In particular, one may be interested in estimating a functional of the form $\theta(F) = \int h(x_1, \cdots, x_m) dF(x_1) \cdots F(x_m)$, $m \leq M$, where F is a marginal distribution of the W^*_{ij}. Depending on the purpose of the investigation, different marginalizations can be employed. The two choices discussed below appear to be most common.

Consider a random pair of indices (I, J) chosen uniformly from the pairs $\{(i, j) : 1 \leq j \leq N_i,\ 1 \leq i \leq M\}$. Let $F_O(w) = EI(W^*_{IJ} \leq w)$, where E denotes expectation w.r.t. the joint distribution of all the random variables involved. In other words, F_O is a marginal distribution of all waiting times irrespective of their cluster membership. A natural estimator of F_O is given by

$$\hat{F}_O(w) = \frac{1}{\sum_{i=1}^{M} N_i} \sum_{i=1}^{M} \sum_{j=1}^{N_i} I(W^*_{ij} \leq w).$$

The corresponding U-statistic is given by

$$\begin{aligned} U_O &= \int h(w_1, \cdots, w_m) d\hat{F}_O(w_1) \cdots \hat{F}_O(w_m) \\ &\approx \frac{1}{\binom{\sum_{i=1}^{M} N_i}{m}} \sum_{\text{all selection of pairs}} h(W^*_{i_1 j_1}, \cdots, W^*_{i_m j_m}). \end{aligned} \tag{3.19}$$

Another marginal distribution that is often considered, when the cluster size is informative, is given by $F_C(w) = EI(W^*_{IJ(I)} \leq w)$, where I follows a discrete uniform on $\{1, \cdots, M\}$, and given $I = i$, J follows a discrete uniform on $\{1, \cdots, N_i\}$. Informative cluster size refers to situations where the size of a cluster is correlated with the distributions of the outcomes in that cluster, often through a hidden factor such as a cluster level random effect (Williamson et al. 2003; Wang et al. 2011; Lorenz et al. 2011; Nevalainen et al. 2014). Note that F_C denotes the distribution of a typical member of a typical cluster in the population. Its natural estimator is given by

$$\hat{F}_C(w) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{N_i} \sum_{j=1}^{N_i} I(W^*_{ij} \leq w).$$

The corresponding U-statistic is given by

$$U_C = \int h(w_1, \cdots, w_m) d\hat{F}_C(w_1) \cdots \hat{F}_C(w_m)$$
$$\approx \frac{1}{\binom{M}{m}} \sum_{1 \le i_1 < \cdots < i_m \le M} N_{i_1}^{-1} \cdots N_{i_m}^{-1} \sum_{j_1(i_1)=1}^{N_{i_1}} \cdots \sum_{j_m(i_m)=1}^{N_{i_m}} h(W^*_{i_1 j_1(i_1)}, \cdots, W^*_{i_m j_m(i_m)}). \tag{3.20}$$

It is not difficult to show that when the cluster size is not informative and the outcomes W^*_{ij} in a cluster are identically distributed, the two marginal distributions F_O and F_C coincide and in this case, both U_O and U_C are valid estimators of the same population functional θ.

In the presence of right censoring, the U-statistics in (3.19) and (3.20) need to be modified to the following approximate U-statistics using the IPCW technique. Thus we define

$$\hat{U}_O(1) \approx \frac{1}{\binom{\sum_{i=1}^M N_i}{m}} \sum_{\text{all selection ofpairs}} \frac{\delta_{i_1 j_1} \cdots \delta_{i_m j_m}}{\hat{S}_c(V_{i_1 j_1} -) \cdots \hat{S}_c(V_{i_m j_m} -)} h(W_{i_1 j_1}, \cdots, W_{i_m j_m}), \tag{3.21}$$

where $\hat{S}_c$, $\delta_{i_k j_k}$, $V_{i_k j_k}$, and $W_{i_k j_k}$ carry the obvious meanings following earlier convention, and

$$\hat{U}_O(2) = \int h(w_1, \cdots, w_m) d\hat{F}_O(w_1) \cdots d\hat{F}_O(w_m), \tag{3.22}$$

where $\hat{F}_O$ will be the Datta–Satten estimator of F_O computed using the pooled sample regardless of the cluster membership. Similarly, we have

$$\hat{U}_C(1) = \frac{1}{\binom{M}{m}} \sum_{1 \le i_1 < \cdots < i_m \le M} N_{i_1}^{-1} \cdots N_{i_m}^{-1}$$
$$\times \sum_{j_1(i_1)=1}^{N_{i_1}} \cdots \sum_{j_1(i_m)=1}^{N_{i_m}} \frac{\delta_{i_1 j_1(i_1)} \cdots \delta_{i_m j_m(i_m)}}{\hat{S}_c(V_{i_1 j_1(i_1)} -) \cdots \hat{S}_c(V_{i_m j_m(i_m)} -)} h(W_{i_1 j_1(i_1)}, \cdots, W_{i_m j_m(i_m)}), \tag{3.23}$$

and

$$\hat{U}_C(2) = \int h(w_1, \cdots, w_m) d\hat{F}_C(w_1) \cdots d\hat{F}_C(w_m), \tag{3.24}$$

where $\hat{F}_C$ will be the Datta–Satten estimator of F_C computed using inverse cluster size weighting in the formulas for the counting and number at risk processes.

Note that to obtain a jackknife estimate of the variance of such U-statistics, the jackknifing should be applied at the cluster level, i.e., an entire (independent) cluster should be deleted at a time in the calculation of jackknife variance. K-sample versions of clustered data U-statistics (3.21)–(3.24) can be obtained in a natural way following the ideas of Sect. 3.4.1.

Note that for the kernel $h(w) = sgn(w)$, one obtains the sign test statistic and with the kernel $h(w_1, w_2) = I(w_1 + w_2 > 0)$, one gets the so-called one-sample Wilcoxon statistic; likewise, a two-sample U-statistic with $h(w_1, w_2) = I(w_1 \leq w_2)$ yields a Wilcoxon rank sum statistic. Clustered data version (U_O without informative cluster size) of these statistics and their multivariate extensions were considered by Larocque et al. (2007) and Nevalainen et al. (2010), although censoring was not considered by these authors.

3.5 An Application to SCI Data

We illustrate inference based on our U-statistic $\hat{U}(2)$ to data from patients undergoing intensive activity-based locomotor training (LT) in the NeuroRecovery Network (NRN) (Harkema et al. 2012). The NRN is a specialized network of treatment centers providing standardized, activity based therapy called locomotor training (LT) for spinal cord injured (SCI) patients. The sample consists of 273 patients with incomplete SCI (a grade of C or D on the International Standards for Neurological Classification of Spinal Cord Injury scale) who were enrolled in the NRN between February 2008 and March 2011. All enrolled patients received standardized locomotor training sessions, as established by NRN protocol, and were evaluated approximately monthly for progress.

Regaining the ability to walk is a frequent goal of rehabilitation for patients with incomplete SCI, and walking speed on the 10 Meter Walk (10MW) test is a frequently used measure of walking capacity for SCI patients. As such, several clinical benchmarks for walking speed have been established: 0.44 m/s represents the minimum walking speed associated with the ability to walk in the community, 0.7 m/s separates those who require assistive walking devices from those who do not, and 1.2 m/s is the approximate speed required to cross a street at a stoplight, a hallmark task of community ambulation (van Hedel and Dietz 2010). Progression through these benchmark speeds can be viewed as a multistate system with 5 states: (1) patient unable to walk (10 MW speed = 0 m/s), (2) patient able to walk no faster than 0.44 m/s, (3) patient able to walk no faster than 0.7 m/s, (4) patient able to walk no faster than 1.2 m/s, and (5) patient able to walk faster than 1.2 m/s (Fig. 3.2). For this illustration, we ignore the potential interval censoring present in this data by assuming that the examination times correspond to actual transition times.

Our interest is in comparing the waiting time in state 2 (walk speed < 0.44 m/s) of the multistate model between two groups of patients defined by a functional classifier measured in the NRN known as the Neuromuscular Recovery Scale (NRS) (Behrman et al. 2012). The NRS measures patient recovery by evaluating

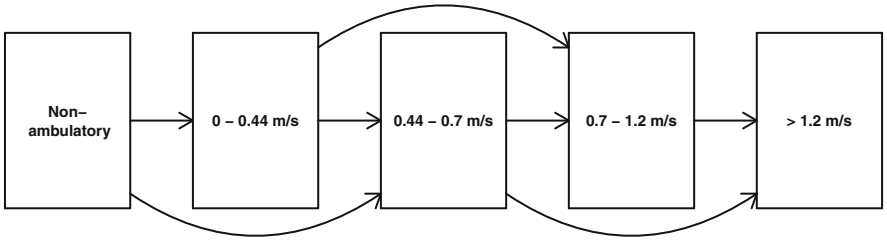

Fig. 3.2 Multistate representation of spinal cord injury patient recovery of walking function

14 functional tasks and scoring performance on these tasks relative to pre-injury capability. The 14 items are aggregated to produce the NRS Phase, which classifies patients into one of four functional groups termed NRS Phase 1, 2, 3, and 4. We are interested in comparing state 2 waiting times between two groups of patients those in NRS Phase 1 and those in NRS Phase 2 or 3 (NRS Phase 4 patients are high functioning and occupy state 2 of the model only rarely).

Censoring in this data set occurred due to exit from the NRN program prior to the attainment of a walking speed of 1.2 m/s. Of the 273 patients in this sample, 122 patients entered state 2 of which 68 were censored, 41 exited to state 3, and 13 exited to state 4. We computed a 2-sample U-statistic $\hat{U}(2)$ with kernel $h(w_{1,i(1,1)}; w_{2,i(2,1)}) = I(w_{1,i(1,1)} > w_{2,i(2,1)})$, as defined in Sect. 3.4.1 to perform a Wilcoxon-type comparison of state 2 waiting times between NRS Phase 1/2 and NRS Phase 3 patients. For these data, $\hat{U}(2) = 0.27$ (as defined in Sect. 4.2) and its jackknife estimated standard error was 0.12 leading to a Z statistic of 2.29 ($p = 0.022$). Therefore, state 2 waiting times for NRS Phase 1 patients were significantly longer than waiting times for Phase 2/3 patients.

3.6 Discussion

This chapter extends the concept of U-statistics from independent and identically distributed complete random variables to those based on incomplete data due to censoring and also when data are clustered. The form of censoring is fairly complex since both state entry and exit times may be right censored and the censoring mechanism may be dependent on external and internal covariates (e.g., past times and stage occupation, etc). Another useful extension was achieved by considering the K-sample case, where many of the familiar statistics fall under this category as well. Collectively, the class of statistics considered in this chapter is very broad and therefore we hope that these statistics will be regarded as useful inferential tools by users.

While we have not studied the detailed asymptotic theory for the statistics proposed here, it is anticipated that the classical linearization tools (e.g., Hoeffding's decomposition Serfling 1980) will apply together with martingale representations

to handle the treatment of censoring by the IPCW technique (Satten and Datta 2001) leading to asymptotic normality of these statistics under appropriate regularity conditions. We have suggested using the jackknife for variance estimation. This together with the normal approximation will suffice for conducting inference.

Other extensions of U-statistics may be possible that will enhance the use of these tools even further. For example, one may consider more complex form of censoring such as the current status and interval censoring. Another extension will be to consider U-statistics using various forms of weights to adjust for certain types of confounding; see Satten and Datta (2014).

Acknowledgements This work was in part supported by the Centers for Disease Control and Prevention and the Christopher and Dana Reeve Foundation (grant/ cooperative agreement U10/ CCU220379). The authors thank the Christopher and Dana Reeve Foundation and all current and past members of the NeuroRecovery Network. We thank two reviewers for constructive comments.

References

Aalen, O.O.: A model for nonparametric regression analysis of counting processes. In: Klonecki, W., Kozek, A., Rosiski, J. (eds.) Mathematical Statistics and Probability. Lecture Notes on Mathematical Statistics and Probability, vol. 2, pp. 1–25. Springer, New York (1980)

Aalen, O.O.: A linear regression model for the analysis of lifetimes. Stat. Med. **8**, 907–925 (1989)

Akritas, M.B.: Empirical processes associated with V-statistics and a class of estimators under random censoring. Ann. Stat. **14**, 619–637 (1986)

Arvesen, J.N.: Jackknifing U-statistics. Ann. Math. Stat. **40**, 2076–2100 (1969)

Behrman, A.L., Ardolino, E., Van Hiel, L., Kern, M., Arkinson, D., Lorenz, D., Harkema, S.: Assessment of functional improvement without compensation reduces variability of outcome measures after human spinal cord injury. Arch. Phys. Med. Rehabil. **93(9)**, 1518–1529 (2012)

Bickel, P.J., Lehmann, E.L.: Descriptive statistics for nonparametric models IV. Spread. In: Jurečková, J. (ed.) Contributions to Statistics, Hajek Memorial Volume, pp. 33–40. Academia, Prague (1979)

Bose, A., Sen, A.: The strong law of large numbers for Kaplan-Meier U- statistics. J. Theor. Probab. **12**, 181–200 (1999)

Bose, A., Sen, A.: Asymptotic distribution of the Kaplan-Meier U-statistics. J. Multivar. Anal. **83**, 84–123 (2002)

Datta, S.: Estimating the mean life time using right censored data. Stat. Methodol. **2**, 65–69 (2005)

Datta, S., Satten, G.A.: Estimation of integrated transition hazards and stage occupation probabilities for non-Markov systems under stage dependent censoring. Biometrics **58**, 792–802 (2002)

Datta, S., Bandyopadhyay, D., Satten, G.A.: Inverse probability of censoring weighted U-statistics for right censored data with applications. Scand. J. Stat. **37**, 680–700 (2010)

Efron, B.: The Jackknife, the Bootstrap, and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia, PA (1982)

Fan, J., Datta , S.: Mann-Whitney tests for comparing sojourn time distributions when the transition times are right censored. Ann. Inst. Stat. Math. **65**, 149–166 (2013)

Gijbels, I., Veraverbeke, N.: Almost sure asymptotic representation for a class of functionals of the Kaplan-Meier estimator. Ann. Stat. **19**, 1457–1470 (1991)

Harkema, S.J., Schmidt-Read, M., Behrman, A.L., Bratta, A., Sisto, S.A., Edgerton, V.R.: Establishing the NeuroRecovery Network: multisite rehabilitation centers that provide activity-based therapies and assessments for neurologic disorders. Arch. Phys. Med. Rehabil. **93**, 1498–1507 (2012)

Hoeffding, W.: A class of statistics with asymptotically normal distribution. Ann. Math. Stat. **19**, 293–325 (1948)

Koul, H., Susarla, V., Van Ryzin, J: Regression analysis of randomly right censored data. Ann. Stat. **9**, 1276–1288 (1981)

Larocque, D., Nevalainen J., Oja, H.: A weighted multivariate sign test for cluster correlated data. Biometrika **94**, 267–283 (2007)

Lee, A.J.: U-Statistics, Theory and Practice. Dekker, New York (1990)

Lorenz, D.J., Datta, S.: A nonparametric analysis of waiting times from a multistate model using a novel linear hazards model approach. Elect. J. Stat. **9**, 419–443 (2015)

Lorenz, D.J., Datta S., Harkema, S.J.: Marginal association measures for clustered data. Stat. Med. **30**, 3181–3191 (2011)

Nevalainen, J. Larocque, D., Oja, H., Pörsti, I: Nonparametric analysis of clustered multivariate data. J. Am. Stat. Assoc. **105**, 864–872 (2010)

Nevalainen, J., Datta, S., Oja, H.: Inference on the marginal distribution of clustered data with informative cluster size. Stat. Pap. **55**, 71–92 (2014)

Randles, R.H., Wolfe, D.A.: Introduction to the Theory of Nonparametric Statistics. Wiley, New York (1979)

Robins, J.M.: Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In: Proceedings of the American Statistical Association, Biopharmaceutical Section, pp. 24–33. Alexandria Soccer Association, Alexandria, VA (1993)

Robins, J.M., Rotnitzky, A.: Recovery of information and adjustment for dependent censoring using surrogate markers. In: Jewell, N., Dietz, K., Farewell, V. (eds.) AIDS Epidemiology—Methodological Issues, pp. 297–331. Birkhauser, Boston (1992)

Satten, G.A., Datta, S.: The Kaplan-Meier Estimator as an inverse-probability- of-censoring weighted average. Amer. Stat. **55**, 207–210 (2001)

Satten, G.A., Datta, S.: Marginal estimation for multi-stage models: waiting time distributions and competing risks analyses. Stat. Med. **21**, 3–19 (2002)

Satten, G.A., Datta, S.: Marginal analyses of multistage data. In: Balakrishnan, N., Rao, C.R. (eds.) Advances in Survival Analysis. Handbook of Statistics, vol. 23, pp. 559–574. Elsevier, Amsterdam (2004)

Satten, G.A., Datta, S.: Multi-sample adjusted U-statistics that account for confounding covariates. Preprint (2014)

Satten, G.A., Datta, S., Robins, J.M.: An estimator for the survival function when data are subject to dependent censoring. Stat. Probab. Lett. **54**, 397–403 (2001)

Schechtman, E., Wang, S.: Jackknifing two-sample statistics. J. Stat. Plan. Inference **119**, 329–340 (2004)

Schisterman, E., Rotnizky, A.: Estimation of the mean of a K-sample U-statistic with missing outcomes and auxiliaries. Biometrika **88**, 713–725 (2001)

Sen, P.K.: Approximation Theorems of Mathematical Statistics. Wiley, New York (1980)

Sen, P.K.: Sequential Nonparametrics: Invariance Principles and Statistical Inference. Wiley, New York (1981)

Stute, W., Wang, J.L.: The strong law under random censorship. Ann. Stat. **21**, 1591–1607 (1993)

Stute, W., Wang, J.L.: Multi-sample U-statistics for censored data. Scand. J. Stat. **20**, 369–374 (1993)

Stute, W.: The bias of Kaplan-Meier integrals. Scand. J. Stat. **21**, 475–484 (1995)

van Hedel, H.J., Dietz, V.: Rehabilitation of locomotion after spinal cord injury. Restor. Neurol. Neurosci. **28**, 123–134 (2010)

Wang, M., Kong, M.K., Datta, S.: Inference for marginal linear models for clustered longitudinal data with potentially informative cluster sizes. Stat. Methods Med. Res. **20**, 347–367 (2011)

Williamson, J.M., Datta, S., Satten, G.A.: Marginal analyses of clustered data when cluster size is informative. Biometrics **59**, 36–42 (2003)

Chapter 4
Nonparametric Location Estimators in the Randomized Complete Block Design

Stefanie Hayoz and Jürg Hüsler

Abstract Several tests for the comparison of different groups in the randomized complete block design exist. However, there is a lack of robust estimators for the location difference between one group and all the others on the original scale. The relative marginal effects are commonly used in this situation, but they are more difficult to interpret and use by less experienced people because of the different scale. In this paper two nonparametric estimators for the comparison of one group against the others in the randomized complete block design will be presented. Theoretical results such as asymptotic normality, consistency, translation invariance, scale preservation, unbiasedness, and median unbiasedness are derived. The finite sample behavior of these estimators is derived by simulations of different scenarios. In addition, possible confidence intervals with these estimators are discussed and their behavior derived also by simulations.

Keywords Confidence intervals • Dependence • Limit distributions • Location estimator • Randomized complete block design • Simulations

4.1 Introduction

Consider the randomized complete block design with k treatments/groups and n blocks. Each treatment is applied exactly once in each block, what leads to $N = k \cdot n$ observations. The blocks could be, e.g., teeth of the same patient, several treatments measured on the same patient, a variable measured at different time points on the same patient, patients matched together by certain variables, animals of the same

S. Hayoz
Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland

Now at Statistics Unit, Swiss Group for Clinical Cancer Research (SAKK), Bern, Switzerland
e-mail: stefanie.hayoz@sakk.ch

J. Hüsler (✉)
Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland
e-mail: juerg.huesler@stat.unibe.ch

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_4

Table 4.1 Data and marginal distributions

	Data			Marginal distribution		
	Treatment/group			Treatment/group		
Block	$j=1$	$\ldots$	$j=k$	$j=1$	$\ldots$	$j=k$
$i=1$	X_{11}	$\ldots$	X_{1k}	F_{11}	$\ldots$	F_{1k}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$i=n$	X_{n1}	$\ldots$	X_{nk}	F_{n1}	$\ldots$	F_{nk}

litter. The sample variables $X_{ij}, i = 1,\ldots,n, j = 1,\ldots,k$, and their marginal distributions F_{ij} are depicted in Table 4.1.

The vectors $\boldsymbol{X}_i = (X_{i1},\ldots,X_{ik})$ are assumed to be independent. We assume that $F_{1j} = \ldots = F_{nj} = F_j, j = 1,\ldots,k$, where the median of F_j is $\alpha_j, j = 1,\ldots,k$. Denote the pairwise differences between the observations by $D_{i\ell,hj} = X_{i\ell} - X_{hj}, i,h = 1,\ldots,n, \ell,j = 1,\ldots,k$.

Three tests for this situation, namely the Friedman test (see, e.g., Friedman 1940), the ANOVA type test (see, e.g., Brunner et al. 2002), and the ANOVA test for repeated measures (see, e.g., Davis 1952), were compared with respect to type I error and power by Hayoz (2006). In addition to the tests, it would be desirable to have a robust estimator for the location difference between one group and all the others.

To describe differences in the distributions, the relative marginal effects

$$p_j = \int H \, \mathrm{d}F_j$$

can be used, where $H = \frac{1}{k}\sum_{j=1}^{k} F_j$. See Brunner et al. (2002) and Konietschke et al. (2010) for further details, also on simultaneous confidence intervals for the relative marginal effects. The relative marginal effects are scale free, which can be an advantage in comparing different variables. However in applications relative marginal effects are often difficult to interpret for people who are not used to them.

Thus, an estimator for the location difference between one group and all the others on the original scale would be desirable. This location difference can be quantified as

$$\lambda_j = \alpha_j - \frac{1}{k-1}\sum_{\substack{\ell=1\\ \ell\neq j}}^{k}\alpha_\ell = \frac{1}{k-1}\sum_{\substack{\ell=1\\ \ell\neq j}}^{k}(\alpha_j - \alpha_\ell).$$

W.l.o.g. the groups can be renumbered which leads to

$$\lambda_1 = \frac{1}{k-1}\sum_{j=2}^{k}\lambda_{1j},$$

where $\lambda_{1j} = \alpha_1 - \alpha_j$. The simplest approach to estimate λ_1 would be based on means,

$$\bar{X}_1 - \frac{1}{k-1}\sum_{j=2}^{k}\bar{X}_j,$$

where $\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, j = 1,\ldots,k$, but this estimator is not robust and not reasonable for skewed distributions. One idea for a nonparametric estimator for λ_1, based on an estimator proposed in Lehmann (1963) for independent groups, is to use the mean of the median of pairwise differences (pd) which leads to the estimator:

$$\hat{\lambda}_1^{pd} = \frac{1}{k-1}\sum_{j=2}^{k}\hat{\lambda}_{1j},$$

where

$$\hat{\lambda}_{1j} = \text{med}(X_{i1} - X_{hj}, i, h = 1,\ldots,n).$$

If the groups are independent, the estimator $\hat{\lambda}_{1j}$ corresponds to the two-sample Wilcoxon statistic. Koul (2002) showed that the Hodges–Lehmann estimator can be derived also as a minimum distance estimator. So the estimator $\hat{\lambda}_1^{pd}$ is a function of minimum distance estimators. A third estimate may be based on the median of all block-wise differences:

$$\hat{\lambda}_1^{block} = \text{med}(X_{i1} - X_{ij}, i = 1,\ldots,n, j = 2,\ldots,k).$$

$\hat{\lambda}_1^{block}$ estimates ξ, the median of $\bar{G}_{\cdot}(x) = \frac{1}{k-1}\sum_{j=2}^{k} G_j(x)$, where G_j denotes the marginal distribution of $X_{i1} - X_{ij}$. Also this estimate may be seen as a minimum distance estimator. Using the approach of Koul (2002) further estimators could be derived. In Hayoz (2006) and Hayoz (2012) several other nonparametric estimators were proposed which can be seen also in this context. In this paper we present and discuss only $\hat{\lambda}_1^{pd}$ and $\hat{\lambda}_1^{block}$ since they proved best with regard to theoretical properties and also performed well in several simulated scenarios. Some theoretical properties will be discussed in Sect. 4.2 and simulation results will be shown in Sect. 4.3. In Sect. 4.4 we will consider the question of confidence intervals.

4.2 Theoretical Results

In this section asymptotic normality, consistency, translation invariance, scale preservation, unbiasedness, and median unbiasedness will be derived for the two estimators proposed in Sect. 4.1.

4.2.1 Mean of the Median of Pairwise Differences

We need to investigate the properties of the estimator $\hat{\lambda}_{1j}$, $j = 2, \ldots, k$. Hodges and Lehmann (1963) and in more detail Lehmann (1963) already proposed the estimator $\hat{\lambda}_{1j}$ for independent groups. We generalize his results for dependent groups assuming only absolutely continuous and differentiable marginal distribution functions. $\hat{\lambda}_{1j}, j = 2, \ldots, k$ is an estimator for λ_{1j}, where λ_{1j} is defined such that $\int F_j(x - \lambda_{1j})\ \mathrm{d}F_1(x) = \frac{1}{2}, j = 2, \ldots, k$. We mentioned that this estimator corresponds to the two-sample Wilcoxon statistic (see, for example, Lehmann 1975, p. 81–91). However, in our case X_{i1} and X_{hj} are only independent if $i \neq h$ and not for all $i, h = 1, \ldots, n$. Hollander et al. (1974) proved asymptotic normality of the two-sample Wilcoxon statistic when the assumption of independence between samples is weakened to allow pairing, concretely for the case where X and Y are random variables with absolutely continuous joint distribution function F and marginal distribution functions F_X and F_Y and $(X_1, Y_1), \ldots, (X_n, Y_n)$, $X_{n+1}, \ldots, X_{n+s}$, $Y_{n+1}, \ldots, Y_{n+t}$ are independent where (X_i, Y_i) is distributed as (X, Y) for $i = 1, \ldots, n$, X_{n+i} is distributed as X for $i = 1, \ldots, s$ and Y_{n+i} is distributed as Y for $i = 1, \ldots, t$. We extend the theorem of Hollander et al. (1974) slightly to prove asymptotic multivariate normality of a vector of pairwise two-sample Wilcoxon statistics. For the sake of simplicity we consider just the case where we have only complete pairs.

Theorem 4.1 *Let $X_1, \ldots X_k$ be random variables with absolutely continuous marginal distribution functions $F_1, \ldots, F_k$. Let $(X_{11}, X_{1j}), \ldots, (X_{n1}, X_{nj})$ be independent where (X_{i1}, X_{ij}) is distributed as (X_1, X_j) for $i = 1, \ldots, n, j = 2, \ldots, k$. Let*

$$T_{X_1,X_j} = \sqrt{\frac{n}{2}} \left[\frac{U_{X_1,X_j}}{n^2} - \mathbf{P}(X_1 < X_j) \right],$$

where $U_{X_1,X_j} = \sum_{i=1}^{n} \sum_{h=1}^{n} 1(X_{i1} < X_{hj})$ and $\mathbf{P}(X < Y) = \int F_1(x)\ \mathrm{d}F_j(x)$. Then as $n \to \infty$,

$$\mathbf{T} = (T_{X_1,X_2}, \ldots, T_{X_1,X_k})$$

converges in distribution to a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\begin{pmatrix} \sigma^2_{T_{X_1,X_2}} & \gamma_{T_{X_1,X_2},T_{X_1,X_3}} & \cdots & \gamma_{T_{X_1,X_2},T_{X_1,X_k}} \\ \gamma_{T_{X_1,X_2},T_{X_1,X_3}} & \sigma^2_{T_{X_1,X_3}} & \cdots & \gamma_{T_{X_1,X_3},T_{X_1,X_k}} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{T_{X_1,X_2},T_{X_1,X_k}} & \gamma_{T_{X_1,X_3},T_{X_1,X_k}} & \cdots & \sigma^2_{T_{X_1,X_k}} \end{pmatrix},$$

where

$$\begin{aligned}\sigma^2_{T_{X_1.X_j}} &= \frac{1}{2}\int [1-F_j(x)]^2\,\mathrm{d}F_1(x) + \frac{1}{2}\int F_1^2(x)\,\mathrm{d}F_j(x)\\ &\quad - \left(\int F_1(x)\,\mathrm{d}F_j(x)\right)^2 - \mathrm{Cov}(F_j(X_{11}), F_1(X_{1j})),\end{aligned} \tag{4.1}$$

$j = 2, \ldots, k$, *and*

$$\begin{aligned}\gamma_{T_{X_1.X_j}, T_{X_1.X_\ell}} &= \frac{1}{2}\,\mathrm{Cov}(1-F_j(X_{11}), 1-F_\ell(X_{11}))\\ &\quad + \frac{1}{2}\,\mathrm{Cov}(1-F_j(X_{11}), F_1(X_{1\ell})) + \frac{1}{2}\,\mathrm{Cov}(F_1(X_{1j}), 1-F_\ell(X_{11}))\\ &\quad + \frac{1}{2}\,\mathrm{Cov}(F_1(X_{1j}), F_1(X_{1\ell})),\end{aligned} \tag{4.2}$$

$j, \ell = 2, \ldots, k, j \neq \ell$.

For the proof see Hayoz (2012). Note that if $F_1 = \ldots = F_k = F$, then (4.1) simplifies to

$$\sigma^2_{T_{X_1.X_j}} = \frac{1}{12} - \mathrm{Cov}(F(X_{11}), F(X_{1j})),$$

$j = 2, \ldots, k$.

With this result we can derive the asymptotic normality of $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_{12}, \ldots, \hat{\lambda}_{1k})$. Assume that

$$F_1, \ldots, F_k \text{ are absolutely continuous with densities } f_j \tag{4.3}$$

Let

$$\boldsymbol{\Sigma}_\Lambda = \begin{pmatrix} \sigma^2_{\hat{\lambda}_{12}} & \gamma_{\hat{\lambda}_{12},\hat{\lambda}_{13}} & \cdots & \gamma_{\hat{\lambda}_{12},\hat{\lambda}_{1k}} \\ \gamma_{\hat{\lambda}_{12},\hat{\lambda}_{13}} & \sigma^2_{\hat{\lambda}_{13}} & \cdots & \gamma_{\hat{\lambda}_{13},\hat{\lambda}_{1k}} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{\hat{\lambda}_{12},\hat{\lambda}_{1k}} & \gamma_{\hat{\lambda}_{13},\hat{\lambda}_{1k}} & \cdots & \sigma^2_{\hat{\lambda}_{1k}} \end{pmatrix},$$

where

$$\sigma^2_{\hat{\lambda}_{1j}} = \sqrt{2}\left\{\frac{1}{2}\int [1-F_1(x+\lambda_{1j})]^2\,\mathrm{d}F_j(x) + \frac{1}{2}\int F_j^2(x-\lambda_{1j})\,\mathrm{d}F_1(x) - \left(\int F_j(x-\lambda_{1j})\,\mathrm{d}F_1(x)\right)^2 - \mathrm{Cov}(F_j(X_{1j}), F_1(X_{11}))\right\} \Big/ \int f_j(x-\lambda_{1j}) f_1(x)\,\mathrm{d}x \tag{4.4}$$

and

$$\gamma_{\hat{\lambda}_{1j},\hat{\lambda}_{1\ell}} = \frac{1}{\sqrt{2}}\left\{\mathrm{Cov}(1-F_j(X_{11}), 1-F_\ell(X_{11})) + \mathrm{Cov}(1-F_j(X_{11}), F_1(X_{1\ell})) + \mathrm{Cov}(F_1(X_{1j}), 1-F_\ell(X_{11})) + \mathrm{Cov}(F_1(X_{1j}), F_1(X_{1\ell}))\right\} \Big/ \sqrt{\int f_j(x-\lambda_{1j}) f_1(x)\,\mathrm{d}x \int f_\ell(x-\xi_{1\ell}) f_1(x)\,\mathrm{d}x}. \tag{4.5}$$

Theorem 4.2 *If (4.3) holds, then as* $n \to \infty$,

$$\hat{\boldsymbol{\Lambda}}_1 = (\sqrt{n}(\hat{\lambda}_{12} - \lambda_{12}), \ldots, \sqrt{n}(\hat{\lambda}_{1k} - \lambda_{1k}))$$

converges in distribution to a multivariate normal distribution with mean $\boldsymbol{0}$ *and covariance matrix* Σ_Λ.

Proof For arbitrary $a_1, \ldots, a_k$

$$\begin{aligned}
&\lim_{n\to\infty} \mathbf{P}(\sqrt{n}(\hat{\lambda}_{1j} - \lambda_{1j}) \le a_j,\ \forall j)\\
&= \lim_{n\to\infty} \mathbf{P}(\mathrm{med}(X_{i1} - X_{hj} - \lambda_{1j} - a_j/\sqrt{n}, i, h = 1, \ldots, n) \le 0,\ \forall j)\\
&= \lim_{n\to\infty} \mathbf{P}(\#\{X_{i1} - \lambda_{1j} - a_j/\sqrt{n} < X_{hj}, i, h = 1, \ldots, n\} \ge \frac{n^2}{2} + O(1),\ \forall j)\\
&= \lim_{n\to\infty} \mathbf{P}(U_{X_j, X_1 - \lambda_{1j} - a_j/\sqrt{n}} \le n^2/2 + O(1),\ \forall j)\\
&= \lim_{n\to\infty} \mathbf{P}\left(\frac{T_{X_j, X_1 - \lambda_{1j} - a_j/\sqrt{n}}}{\sigma_{T_{X_j, X_1 - \lambda_{1j} - a_j/\sqrt{n}}}} \le \sqrt{\frac{n}{2}}\frac{1/2 - \mathbf{P}(X_j < X_1 - \lambda_{1j} - a_j/\sqrt{n}) + O\left(1/n^2\right)}{\sigma_{T_{X_j, X_1 - \lambda_{1j} - a_j/\sqrt{n}}}},\ \forall j\right).
\end{aligned}$$

Let us consider the term $\sigma^2_{T_{X_j,X_1-\lambda_{1j}-a_j/\sqrt{n}}}$. For each j by (4.1)

$$\begin{aligned}
&\lim_{n\to\infty} \sigma^2_{T_{X_j,X_1-\lambda_{1j}-a_j/\sqrt{n}}} \\
&= \lim_{n\to\infty} \frac{1}{2} \int [1 - F_1(x+\lambda_{1j}+a_j/\sqrt{n})]^2 \, \mathrm{d}F_j(x) + \lim_{n\to\infty} \frac{1}{2} \int F_j^2(x-\lambda_{1j}-a_j/\sqrt{n}) \, \mathrm{d}F_1(x) \\
&\quad - \lim_{n\to\infty} \left(\int F_j(x-\lambda_{1j}-a_j/\sqrt{n})) \, \mathrm{d}F_1(x) \right)^2 - \lim_{n\to\infty} \mathrm{Cov}(F_j(X_j), F_1(X_1))
\end{aligned}$$

With the help of the dominated convergence theorem (see, e.g., Forster 1999) the first term of $\sigma^2_{T_{X_j,X_1-\lambda_{1j}-a_j/\sqrt{n}}}$ tends to

$$\frac{1}{2} \int [1 - F_1(x + \lambda_{1j})]^2 \, \mathrm{d}F_j(x),$$

since

$$\lim_{n\to\infty} [1 - F_1(x+\lambda_{1j}+a_j/\sqrt{n})]^2 = [1 - F_1(x + \lambda_{1j})]^2.$$

Using similar arguments for the second and third term, we get for each j

$$\begin{aligned}
&\lim_{n\to\infty} \sigma^2_{T_{X_j,X_1-\lambda_{1j}-a_j/\sqrt{n}}} \\
&= \frac{1}{2} \int [1 - F_1(x + \lambda_{1j})]^2 \, \mathrm{d}F_j(x) + \frac{1}{2} \int F_j^2(x - \lambda_{1j}) \, \mathrm{d}F_1(x) \\
&\quad - \left(\int F_j(x - \lambda_{1j}) \, \mathrm{d}F_1(x) \right)^2 - \mathrm{Cov}(F_j(X_j), F_1(X_1)) \\
&= \sigma^2_{\hat{\lambda}_{1j}} \frac{1}{\sqrt{2}} \int f_j(x - \lambda_{1j}) f_1(x) \, \mathrm{d}x.
\end{aligned} \tag{4.6}$$

Now we consider

$$\begin{aligned}
&\lim_{n\to\infty} \sqrt{n} \, [\mathbf{P}(X_j < X_1 - \lambda_{1j}) - \mathbf{P}(X_j < X_1 - \lambda_{1j} - a_j/\sqrt{n})] \\
&= \lim_{n\to\infty} \sqrt{n} \int [F_j(x - \lambda_{1j}) - F_j(x - \lambda_{1j} - a_j/\sqrt{n})] \, \mathrm{d}F_1(x).
\end{aligned} \tag{4.7}$$

Using a Taylor series expansion and the dominated convergence theorem we get for each j

$$\begin{aligned}
&\lim_{n\to\infty} \sqrt{n} \int [F_j(x-\lambda_{1j}) - F_j(x-\lambda_{1j}-a_j/\sqrt{n})]\, \mathrm{d}F_1(x) \\
&= \lim_{n\to\infty} \sqrt{n} \int \left[F_j(x-\lambda_{1j}) - \left(F_j(x-\lambda_{1j}) - \frac{a_j}{\sqrt{n}} f_j(x-\lambda_{1j}) - o\left(\frac{a_j}{\sqrt{n}}\right)\right)\right] \mathrm{d}F_1(x) \\
&= a_j \int f_j(x-\lambda_{1j}) f_1(x)\, \mathrm{d}x.
\end{aligned} \tag{4.8}$$

Note that by definition of λ_{1j} for each j

$$\mathbf{P}(X_j < X_1 - \lambda_{1j}) = \int F_j(x-\lambda_{1j})\, \mathrm{d}F_1(x) = \frac{1}{2}. \tag{4.9}$$

Now for each j

$$\begin{aligned}
&\sqrt{\frac{n}{2}} \frac{1/2 - \mathbf{P}(X_j < X_1 - \lambda_{1j} - a_j/\sqrt{n})}{\sigma_{T_{X_j, X_1-\lambda_{1j}-a_j/\sqrt{n}}}} \\
&= \sqrt{\frac{n}{2}} \frac{1/2 - \mathbf{P}(X_j < X_1 - \lambda_{1j} - a_j/\sqrt{n})}{\sigma_{1j}\frac{1}{\sqrt{2}} \int f_j(x-\lambda_{1j}) f_1(x)\, \mathrm{d}x} \frac{\sigma_{1j}\frac{1}{\sqrt{2}} \int f_j(x-\lambda_{1j}) f_1(x)\, \mathrm{d}x}{\sigma_{T_{X_j, X_1-\lambda_{1j}-a_j/\sqrt{n}}}} \\
&= \underbrace{\sqrt{\frac{n}{2}} \frac{\mathbf{P}(X_j < X_1 - \lambda_{1j}) - \mathbf{P}(X_j < X_1 - \lambda_{1j} - a_j/\sqrt{n})}{\sigma_{1j}\frac{1}{\sqrt{2}} \int f_j(x-\lambda_{1j}) f_1(x)\, \mathrm{d}x}}_{\to a_j/\sigma_{\hat{\lambda}_{1j}} \text{ by (4.7) and (4.8)}} \underbrace{\frac{\sigma_{1j}\frac{1}{\sqrt{2}} \int f_j(x-\lambda_{1j}) f_1(x)\, \mathrm{d}x}{\sigma_{T_{X_j, X_1-\lambda_{1j}-a_j/\sqrt{n}}}}}_{\to 1 \text{ by (4.6)}} \\
&\quad + \underbrace{\sqrt{\frac{n}{2}} \frac{1/2 - \mathbf{P}(X_j < X_1 - \lambda_{1j})}{\sigma_{1j} \int f_j(x-\lambda_{1j}) f_1(x)\, \mathrm{d}x}}_{=0 \text{ by (4.9)}} \underbrace{\frac{\sigma_{1j}\frac{1}{\sqrt{2}} \int f_j(x-\lambda_{1j}) f_1(x)\, \mathrm{d}x}{\sigma_{T_{X_j, X_1-\lambda_{1j}-a_j/\sqrt{n}}}}}_{\to 1 \text{ by (4.6)}}.
\end{aligned}$$

Thus

$$\lim_{n\to\infty} \mathbf{P}(\sqrt{n}(\hat{\lambda}_{1j} - \lambda_{1j}) \le a_j,\ \forall j) = \lim_{n\to\infty} \mathbf{P}\left(\frac{T_{X_j, X_1-\lambda_{1j}-a_j/\sqrt{n}}}{\sigma_{T_{X_j, X_1-\lambda_{1j}-a_j/\sqrt{n}}}} \le \frac{a_j}{\sigma_{\hat{\lambda}_{1j}}},\ \forall j \right).$$

By Theorem 4.1 this is equal to a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ_Λ, as for $j = 2, \ldots, k$,

$$\lim_{n\to\infty} \mathrm{Var}\left(\frac{T_{X_j-\lambda_{1j}-a_j/\sqrt{n}, X_1}}{\sigma_{T_{X_j-\lambda_{1j}-a_j/\sqrt{n}, X_1}}} \sigma_{\hat{\lambda}_{1j}} \right) = \sigma^2_{\hat{\lambda}_{1j}}$$

by (4.6) and (4.4) and for $j, \ell = 2, \ldots, k, j \neq \ell$,

$$\lim_{n\to\infty} \mathrm{Cov}\left(\frac{T_{X_j, X_1 - \lambda_{1j} - a_j/\sqrt{n}}}{\sigma_{T_{X_j, X_1 - \lambda_{1j} - a_j/\sqrt{n}}}} \sigma_{\hat{\lambda}_{1j}}, \frac{T_{X_\ell, X_1 - \xi_{1\ell} - a_\ell/\sqrt{n}}}{\sigma_{T_{X_\ell, X_1 - \xi_{1j\ell} - a_\ell/\sqrt{n}}}} \sigma_{\hat{\lambda}_{1\ell}} \right)$$

$$= 2 \frac{\gamma_{T_{X_j, X_1 - \lambda_{1j}}, T_{X_\ell, X_1 - \xi_{1\ell}}}}{\sqrt{\int f_j(x - \lambda_{1j}) f_1(x)\, \mathrm{d}x \int f_\ell(x - \xi_{1\ell}) f_1(x)\, \mathrm{d}x}})$$

$$= \gamma_{\hat{\lambda}_{1j}, \hat{\lambda}_{1\ell}}$$

using (4.6), (4.2), and (4.5) . □

Note that if $F_1 = \ldots = F_k = F$ or if

$$F_j(x) = F(x - \alpha_j), \forall j, \tag{4.10}$$

then $\sigma^2_{\hat{\lambda}_{1j}}$ in (4.4) can be simplified to

$$\sigma^2_{\hat{\lambda}_{1j}} = \sqrt{2} \left\{ \frac{1}{12} - \mathrm{Cov}(F(X_{1j} - \alpha_j), F(X_{11} - \alpha_1)) \right\} \Big/ \int f^2(x)\, \mathrm{d}x, \tag{4.11}$$

$j = 2, \ldots, k$. Now, the estimator $\hat{\lambda}_{1j}$ has some more properties (see Hayoz 2012 for proofs). Obviously it is translation invariant and scale preserving,

Corollary 4.1 *If (4.3) holds, then $\hat{\lambda}_{1j}$ is a consistent, asymptotically unbiased, and asymptotically median unbiased estimator for λ_{1j}.*

Let $X_{ij} \sim F_j$ with median α_j and $(X_{i1}, X_{ij}) \sim F_{1j}, \forall i, j$. If we assume that

$$F_{1j} \text{ is symmetric about } (\alpha_1, \alpha_j), \forall j, \tag{4.12}$$

i.e. $\mathbf{P}(X_{i1} \leq \alpha_1 + u, X_{ij} \leq \alpha_j + v) = \mathbf{P}(X_{i1} \geq \alpha_1 - u, X_{ij} \geq \alpha_j - v), \forall i, j$, or that

$$X_{i1} - \alpha_1 \text{ and } X_{ij} - \alpha_j \text{ are exchangeable } \forall i, j, \tag{4.13}$$

i.e. $F_{1j}(u + \alpha_1, v + \alpha_j) = F_{1j}(v + \alpha_j, u + \alpha_1), \forall j$, then the distribution of $\hat{\lambda}_{1j}$ is symmetric about $\alpha_1 - \alpha_j$ and thus we can state unbiasedness of $\hat{\lambda}_{1j}$.

Corollary 4.2 *If (4.3) with (4.12) or (4.13) holds, then*

(i) $\hat{\lambda}_{1j}$ is median unbiased estimator for $\alpha_1 - \alpha_j$.
(ii) $\hat{\lambda}_{1j}$ is an unbiased estimator for $\alpha_1 - \alpha_j$, if $\mathrm{E}(\hat{\lambda}_{1j})$ exists.

Note that if $\mathrm{E}(X_{ij})$ and $\mathrm{Var}(X_{ij})$ exist for $i = 1, \cdots, n, j = 1, \cdots, k$, then it can be shown that $\hat{\lambda}_{1j}$ is bounded and thus $\mathrm{E}(\hat{\lambda}_{1j})$ exists.

With these results we derive the properties of the estimator $\hat{\lambda}_1^{pd}$ which is obviously also translation invariant and scale preserving. If (4.12) or (4.13) holds, $\hat{\lambda}_1^{pd}$ estimates

$$\xi_1 = \alpha_1 - \frac{1}{k-1}\sum_{j=2}^{k}\alpha_j,$$

as the distribution of $\hat{\lambda}_{1j}$ is symmetric about $\alpha_1 - \alpha_j$ and thus

Theorem 4.3 *If (4.3) holds, then as* $n \to \infty$,

$$\sqrt{n}\left(\hat{\lambda}_1^{pd} - \lambda_1\right)$$

converges in distribution to a normal distribution with mean 0 and variance

$$\sigma^2_{\hat{\lambda}_1^{pd}} = \frac{1}{(k-1)^2}\sum_{j=2}^{k}\sigma^2_{\hat{\lambda}_{1j}} + \frac{2}{(k-1)^2}\sum_{j=2}^{k-1}\sum_{\ell=j+1}^{k}\gamma_{\hat{\lambda}_{1j},\hat{\lambda}_{1\ell}}. \tag{4.14}$$

Proof We can write

$$\sqrt{n}\left(\hat{\lambda}_1^{pd} - \lambda_1\right) = \sqrt{n}\left(\frac{1}{k-1}\sum_{j=2}^{k}\hat{\lambda}_{1j} - \frac{1}{k-1}\sum_{j=2}^{k}\lambda_{1j}\right)$$

$$= \frac{1}{k-1}\sum_{j=2}^{k}\sqrt{n}(\hat{\lambda}_{1j} - \lambda_{1j})$$

By Theorem 4.2, $\hat{\boldsymbol{\Lambda}}_1$ has an asymptotic multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ_Λ. Thus $\sqrt{n}(\hat{\lambda}_1^{pd} - \lambda_1)$ has an asymptotic normal distribution with mean 0 and variance

$$\frac{1}{(k-1)^2}\sum_{j=2}^{k}\sigma^2_{\hat{\lambda}_{1j}} + \frac{2}{(k-1)^2}\sum_{j=2}^{k-1}\sum_{\ell=j+1}^{k}\gamma_{\hat{\lambda}_{1j},\hat{\lambda}_{1\ell}}.$$

□

Using Corollaries 4.1 and 4.2 we get (see Hayoz 2012 for proofs):

Corollary 4.3 *If (4.3) holds, then* $\hat{\lambda}_1^{pd}$ *is a consistent, asymptotically unbiased, and asymptotically median unbiased estimator for* λ_1.

Corollary 4.4 *If (4.3) with (4.12) or (4.13) holds, then* $\hat{\lambda}_1^{pd}$ *is an unbiased estimator for* ξ_1 *if* $E(\hat{\lambda}_1^{pd})$ *exists.*

Again if $\mathrm{E}(X_{ij})$ and $\mathrm{Var}(X_{ij})$ exist for $i = 1, \cdots, n, j = 1, \cdots, k$, then it can be shown that $\hat{\lambda}_1^{pd}$ is bounded and thus $E(\hat{\lambda}_1^{pd})$ exists. Let $(X_{i1}, \ldots, X_{ik}) \sim F_{1\ldots k}, \forall i$, and assume that

$$F_{1,\ldots,k} \text{ is symmetric about } (\alpha_1, \ldots, \alpha_k), \tag{4.15}$$

i.e.

$$\begin{aligned} &\mathbf{P}(X_{i1} \le \alpha_1 + u_1, \ldots, X_{ik} \le \alpha_k + u_k) \\ &= \mathbf{P}(X_{i1} \ge \alpha_1 - u_1, \ldots, X_{ik} \ge \alpha_k - u_k), \forall i. \end{aligned}$$

Corollary 4.5 *If (4.3) and (4.15) hold, then $\hat{\lambda}_1^{pd}$ is a median unbiased estimator for λ_1.*

More generally we could use weighted means which leads to estimators of the form

$$\hat{\lambda}_1^{w} = \sum_{j=2}^{k} \omega_j \hat{\lambda}_{1j}, \tag{4.16}$$

where

$$\sum_{j=2}^{k} \omega_j = 1.$$

The more general estimator $\hat{\lambda}_1^{w}$ could be useful, for example, if the groups have a fixed order (e.g., longitudinal data or several teeth per person) and we want to give more weight to the differences of groups that lie close together. It can easily be shown that all the above theorems and corollaries hold for $\hat{\lambda}_1^{w}$ as well.

4.2.2 Median of Blockwise Differences

In this section we will present some properties of the estimator $\hat{\lambda}_1^{block}$. We consider only the block-wise differences $D_{i;j} := X_{i1} - X_{ij}$. Denote the distribution function of $D_{i;j}$ by G_j and the joint distribution function of $(D_{i;j}, D_{i;\ell})$ by $G_{j,\ell}$ for $i = 1, \ldots, n, j, \ell = 2, \ldots, k, j \le \ell$. Let ξ be the median of $\bar{G}_{\cdot}(x) = \frac{1}{k-1} \sum_{j=2}^{k} G_j(x)$ and assume that for each $j \le k$

$$G_j \text{ is continuously differentiable in the neighborhood of } \xi \tag{4.17}$$

with

$$0 < \inf_j g_j(\xi) \leq \sup_j g_j(\xi) < \infty, \tag{4.18}$$

where $g_j(x) = \frac{d}{dx}G_j(x)$. We let $g_\cdot(x) = \frac{d}{dx}\bar{G}_\cdot(x) = \frac{1}{k-1}\sum_{j=2}^{k} g_j(x)$. Note that $D_{i;j}$ and $D_{h;\ell}$ are independent for $i \neq h$, $j, \ell = 1, \ldots, k$. With the notation

$$D^{((k-1)(i-1)+j-1)} := D_{i;j} \tag{4.19}$$

it is seen immediately that $D^{(1)}, \ldots, D^{(N)}$, $N = n(k-1)$ form a $(k-1)$-dependent random sequence, that is, the random vectors $(D^{(1)}, \ldots, D^{(\ell)})$ and $(D^{(\ell+u)}, \ldots)$ are stochastically independent if $u > k-1$. The main theorem (Theorem 2.1 on p. 1725) of Sen (1968) states asymptotic normality of sample quantiles for m-dependent processes:

Theorem 4.4 (Theorem 2.1 of Sen 1968) *Let $X_\ell, \ell \geq 1$ be an m-dependent process. Denote the distribution function of X_ℓ by F_ℓ and the joint distribution function of $(X_\ell, X_{\ell+u})$ by $F_{\ell,\ell+u}$ for $\ell = 1, \ldots, N$ and $u = 1, \ldots, m$. If the sample is $X_1, \ldots, X_N$, let $\tilde{X}_N = \text{med}(X_\ell, \ell = 1, \ldots N)$, $\bar{F}_N(x) = \frac{1}{N}\sum_{\ell=1}^{N} F_\ell(x)$ and ξ_N be defined such that $\bar{F}_N(\xi_N) = 1/2$. If $\sup_N |\xi_N| < \infty$,*

$$F_\ell(x) \text{ is absolutely continuous in the neighborhood of } \xi_N, \forall \ell, \tag{4.20}$$

$$f_\ell(x) = \frac{d}{dx}F_\ell(x) \text{ is continuous in some neighborhood of } \xi_\ell, \forall \ell, \tag{4.21}$$

with

$$0 < \liminf_{n\to\infty} f_N(\xi_N) \leq \limsup_{N\to\infty} f_N(\xi_N) < \infty, \tag{4.22}$$

(which implies that $0 < \inf_n \bar{f}_N(\xi_N) \leq \sup_N \bar{f}_N(\xi_N) < \infty$, where $\bar{f}_N(x) = \frac{1}{N}\sum_{\ell=1}^{N} f_\ell(x) = \frac{d}{dx}\bar{F}_N(x)$), and

$$\inf_n v_{N,m}^2 > 0, \tag{4.23}$$

with

$$v_{N,m}^2 = \frac{1}{4} - \frac{1}{N}\sum_{\ell=1}^{N}(F_\ell(\xi_N) - 1/2)^2 + \frac{2}{N}\sum_{\ell=2}^{N}\sum_{u=1}^{m-\ell}(F_{\ell,\ell+u}(\xi_N) - F_\ell(\xi_N)F_{\ell+u}(\xi_N)),$$

then as $N \to \infty$

$$\sqrt{N}(\tilde{X}_N - \xi_N)$$

converges in distribution to a normal distribution with mean 0 and variance $\nu_{N,m}^2/\bar{f}_N^2(\xi_N)$.

Applying Theorem 4.4 we derive the asymptotic normality of $\hat{\lambda}_1^{block}$ which is obviously translation invariant and scale preserving. Let

$$\nu^2 = \frac{1}{4} - \frac{1}{k-1}\sum_{j=2}^{k}(G_j(\xi) - 1/2)^2 + \frac{2}{k-1}\sum_{j=2}^{k-1}\sum_{u=1}^{k-j}(G_{j,j+u}(\xi) - G_j(\xi)G_{j+u}(\xi)),$$

$$\sigma^2_{\hat{\lambda}_1^{block}} = \nu^2/\bar{g}_\cdot^2(\xi) \tag{4.24}$$

and assume that

$$\nu^2 > 0. \tag{4.25}$$

Theorem 4.5 *If (4.17), (4.18) and (4.25) holds, then as* $n \to \infty$,

$$\sqrt{n(k-1)}(\hat{\lambda}_1^{block} - \xi)$$

converges in distribution to a normal distribution with mean 0 and variance $\sigma^2_{\hat{\lambda}_1^{block}}$.

Proof As $D^{(1)}, \ldots D^{(N)}$, by (4.19) from a $(k-1)$-dependent process we use Theorem 4.4 for $D^{(\ell)}$ with $m = k-1$, $N = n(k-1)$ As the distribution function of $D_{i;j}$ is G_j and the joint distribution function of $(D_{i;j}, D_{i;j+u})$ is $G_{j,j+u}$ for $i = 1, \ldots, n, j = 2, \ldots, k, u = 1, \ldots, k-j$, we have $F_\ell = G_j$ and $F_{\ell,\ell+u} = G_{j,j+u}$, where $\ell = (k-1)(i-1)+j-1$. Thus $\bar{F}_n = \bar{G}_\cdot$ and $\xi_n = \xi$ do not depend on n. Now the conditions (4.20), (4.21), and (4.22) can be rewritten as (4.17) and (4.18). Note that $\nu_n^2 = \nu^2$ does not depend on n either, thus condition (4.23) reduces to (4.25). □

Corollary 4.6 *If (4.17), (4.18), and (4.25) hold,* $\hat{\lambda}_1^{block}$ *is a consistent, asymptotically unbiased, and asymptotically median unbiased estimator for* ξ.

Let

$$\alpha_2 = \ldots = \alpha_k. \tag{4.26}$$

This condition holds, for example, in multiple matching where several observations are matched to one control.

Corollary 4.7 *If (4.3), (4.15), and (4.26) hold, then*

(i) $\hat{\lambda}_1^{block}$ *is median unbiased estimator for* $\alpha_1 - \alpha_j$.
(ii) $\hat{\lambda}_1^{block}$ *is an unbiased estimator for* $\alpha_1 - \alpha_j$ *if* $\mathrm{E}(\hat{\lambda}_1^{block})$ *exists.*

Note that if $\mathrm{E}(X_{ij})$ and $\mathrm{Var}(X_{ij})$ exist for $i = 1, \cdots, n, j = 1, \cdots, k$, then it can be shown that $\hat{\lambda}_1^{block}$ is bounded and thus $\mathrm{E}(\hat{\lambda}_1^{block})$ exists. Assume that if (4.26) holds, the following symmetry relation holds for $\xi = \alpha_1 - \alpha_k$ and any $(v_2, \ldots, v_k)$ with $v_j \in \{-1, 1\}$:

$$\begin{aligned} &\mathbf{P}(v_2(D_{i;2} - \xi) \leq 0, \ldots, v_k(D_{i;k} - \xi) \leq 0) \\ &= \mathbf{P}(-v_2(D_{i;2} - \xi) \leq 0, \ldots, -v_k(D_{i;k} - \xi) \leq 0). \end{aligned} \tag{4.27}$$

Let $(D_{i;2}, \ldots, D_{i;k}) \sim G_{1\ldots k}, \forall i$, and assume that

$$D_{i;2}, \ldots, D_{i;k} \text{ are exchangeable } \forall i, \tag{4.28}$$

i.e. $G_{1\ldots k}(u_2, \ldots, u_k) = G_{1\ldots k}(\pi(u_2), \ldots, \pi(u_k))$ for any permutation π of $u_2, \ldots, u_k$.

Corollary 4.8 *If (4.3), (4.27), and (4.28) hold, then $\hat{\lambda}_1^{block}$ is a median unbiased estimator for ξ if $n(k-1)$ is odd and an approximately median unbiased estimator for ξ if $n(k-1)$ even.*

4.3 Simulations

With simulations we analyze and compare the behavior of the estimators of Sect. 4.1 in different situations. For the dependence between $X_{i1}, \ldots, X_{ik}$ of the block i, we assume compound symmetry (CS), first order autoregressive (AR(1)) and first order moving average (MA(1)) as covariance structures. The covariance structures CS, AR(1) and MA(1) were modeled such that $\sigma^2 = 1$ and $\rho = \mathbf{C}\mathrm{orr}(X_{ij}, X_{ij+1}) = 0.2, 0.5, 0.8$ (for MA(1) $\rho = 0.8$ is not possible). The dependence was assumed to be the same for each block. To be able to easily simulate data with different covariance structures and different marginal distributions, copulas were used. For our simulation a t-copula with 3 degrees of freedom simulated with the R package copula was used (see, e.g., Yan 2007). Several continuous symmetric and asymmetric underlying distributions were simulated, including some heavy-tailed distributions to create outliers. Further simulations were run in which the data obtained from continuous distributions were rounded to create tied observations. Several block sizes and group numbers were considered. In this work the results for the standard normal distributions and standard lognormal distribution are shown with $n = 10$ blocks and $k = 4$ groups. For each situation $n_{\mathrm{sim}} = 10'000$ samples were simulated. To show the simulation error of the bias, the intervals $[\mathrm{mean}_{\mathrm{sim}} - 1.96\, \mathrm{s}_{\mathrm{sim}} / n_{\mathrm{sim}}, \mathrm{mean}_{\mathrm{sim}} + 1.96\, \mathrm{s}_{\mathrm{sim}} / n_{\mathrm{sim}}]$, where $\mathrm{mean}_{\mathrm{sim}}$ is the simulated mean and $\mathrm{s}_{\mathrm{sim}}$ the simulated standard deviation, were derived. To show the simulation error of the median bias, confidence intervals based on interpolated order statistics introduced by Hettmansperger and Sheather (1986) were calculated. The

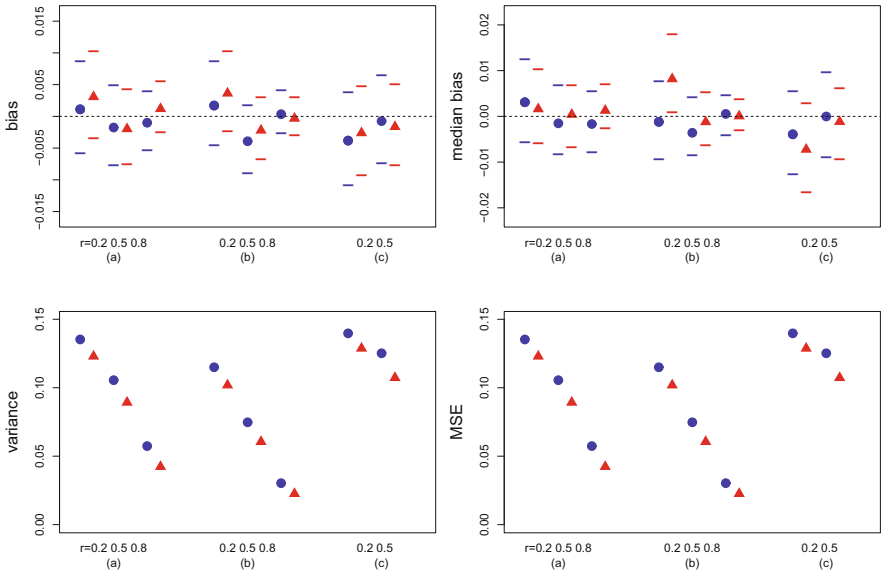

Fig. 4.1 Simulation results of the bias and median bias with confidence intervals as well as variance and MSE of the estimators $\hat{\lambda}_1^{block}$ (*red filled triangle*) and $\hat{\lambda}_1^{pd}$ (*blue filled circle*) with $F_1 = \ldots = F_k, k = 4$ and $n = 10$ with an underlying $\mathcal{N}(0, 1)$ distribution and CS (*a*), AR(1) (*b*) and MA(1) (*c*) covariance structures

simulated biases and median biases with confidence intervals, variances, and MSEs under $F_1 = \ldots = F_k$ are depicted in Figs. 4.1 and 4.2.

Furthermore, the impact of the group location parameters was investigated assuming the semi-parametric location model $F_j = F(x - \alpha_j), \forall j$. The results for $\alpha_2 = \ldots = \alpha_k$ are not shown here as the results in thus situation are the same as under $F_1 = \ldots = F_k$. The simulated biases and median biases with confidence intervals, variances and MSEs with $\alpha_1 = 0, \alpha_2 = 1, \alpha_3 = 2, \alpha_4 = 10$ are depicted in Figs. 4.3 and 4.4. In this situation $\hat{\lambda}_1^{pd}$ estimates $\frac{1}{k-1}\sum_{j=2}^{k}(\alpha_1 - \alpha_j) = 2.33$ and $\hat{\lambda}_1^{block}$ estimates ξ, the median of $\bar{G}_\cdot(x) = \frac{1}{k-1}\sum_{j=2}^{k} G_j(x)$, which depends on the covariance structure and correlation.

Generally the higher ρ, the smaller the bias, the median bias, the variance and thus the MSE. With underlying CS structures we get the smallest variances and MSEs, with underlying MA(1) structures the highest.

If $F_1 = \ldots = F_k$ both considered estimators are unbiased and median unbiased for symmetric marginal distributions. $\hat{\lambda}_1^{block}$ is slightly biased for asymmetric marginal distributions, but always median unbiased. $\hat{\lambda}_1^{pd}$, on the other hand, is unbiased for asymmetric marginal distributions, but not median unbiased. These simulations results are consistent with the theoretical results of Corollary 4.4, Corollary 4.5, Corollary 4.7, and Corollary 4.8. For all marginal distributions $\hat{\lambda}_1^{block}$ has the smallest variance and MSE.

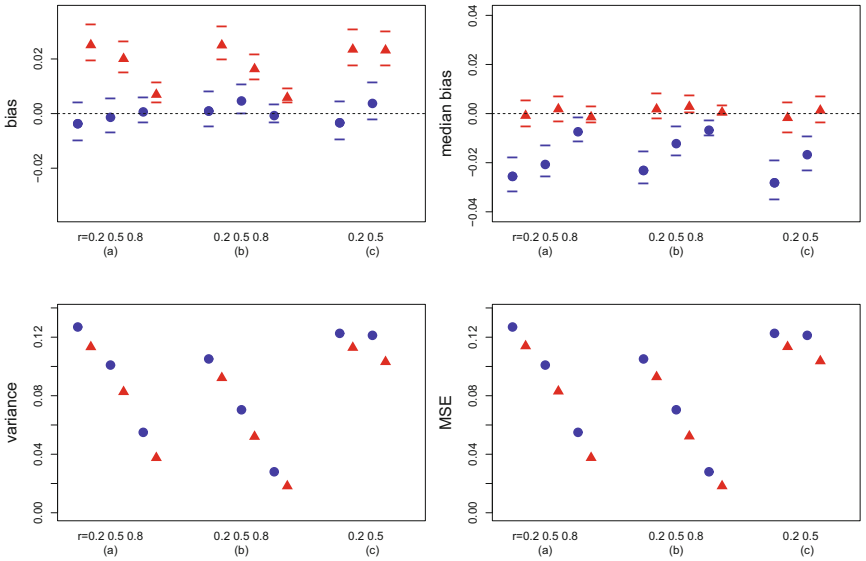

Fig. 4.2 Simulation results of the bias and median bias with confidence intervals as well as variance and MSE of the estimators $\hat{\lambda}_1^{block}$ (*red filled triangle*) and $\hat{\lambda}_1^{pd}$ (*blue filled circle*) with $F_1 = \ldots = F_k, k = 4$ and $n = 10$ with an underlying $\log\mathcal{N}(0,1)$ distribution and CS (*a*), AR(1) (*b*) and MA(1) (*c*) covariance structures

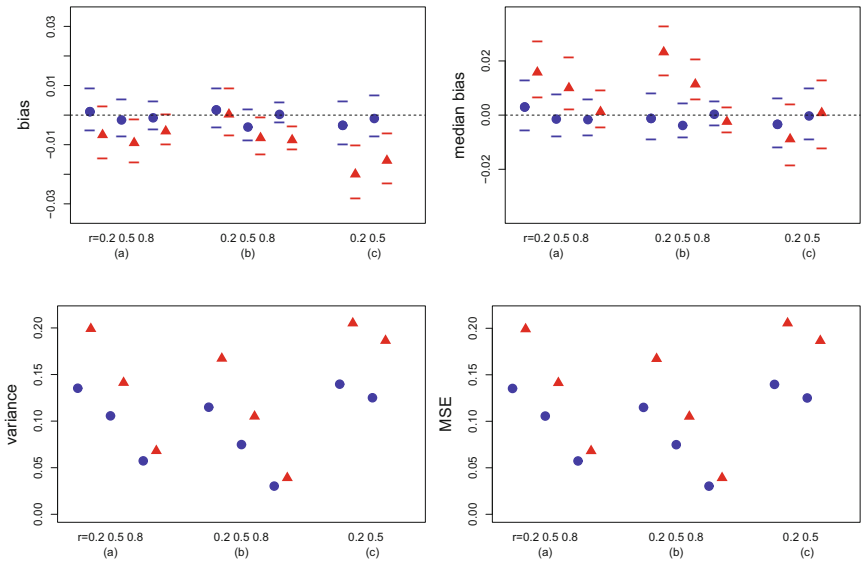

Fig. 4.3 Simulation results of the bias and median bias with confidence intervals as well as variance and MSE of the estimators $\hat{\lambda}_1^{block}$ (*red filled triangle*) and $\hat{\lambda}_1^{pd}$ (*blue filled circle*) with $F_j = F(x-\alpha_j), \forall j, k = 4$ and $n = 10$ with an underlying $\mathcal{N}(0,1)$ distribution and CS (*a*), AR(1) (*b*) and MA(1) (*c*) covariance structures

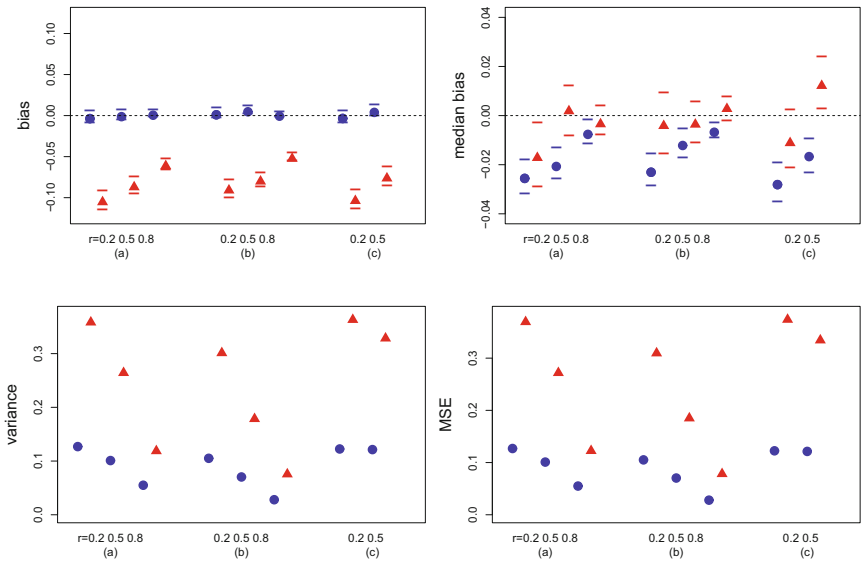

Fig. 4.4 Simulation results of the bias and median bias with confidence intervals as well as variance and MSE of the estimators $\hat{\lambda}_1^{block}$ (*red filled triangle*) and $\hat{\lambda}_1^{pd}$ (*blue filled circle*) with $F_j = F(x - \alpha_j), \forall j, k = 4$ and $n = 10$ with an underlying $\log\mathcal{N}(0, 1)$ distribution and CS (*a*), AR(1) (*b*) and MA(1) (*c*) covariance structures

If $\alpha_1 = 0, \alpha_2 = 1, \alpha_3 = 2, \alpha_4 = 10$ $\hat{\lambda}_1^{block}$ is biased, median biased and has a high variance and MSE for all marginal distributions whereas $\hat{\lambda}_1^{pd}$ is unbiased and for symmetric marginal distributions is also median unbiased.

Further scenarios were considered for which the results can be found in Hayoz (2012) and are only shortly described here. With increasing n the bias, median bias, variance and thus also the MSE decrease. With higher k the variance and MSE get smaller, the bias and median bias stay more or less the same. In scenarios where the variances are not the same in each group both estimators are unbiased and median unbiased with underlying symmetric distributions, whereas with underlying asymmetric distributions both estimators are biased, $\hat{\lambda}_1^{pd}$ is median unbiased, whereas $\hat{\lambda}_1^{block}$ is slightly median biased for small n. When the values from continuous distributions were rounded to create tied observations, both considered estimators showed the same behavior like for the continuous distributions, also with similar variances. Both estimators were not affected by outliers. For comparison the usual mean difference was included in the simulations. However it had higher variances and MSEs than the two nonparametric estimators and is thus not a good alternative.

In conclusion $\hat{\lambda}_1^{block}$ is the best estimator if $F_2 = \ldots = F_k$, i.e. under the nullhypothesis or in multiple matching. If not all groups have the same mean, $\hat{\lambda}_1^{block}$

should not be used, $\hat{\lambda}_1^{pd}$ is the better choice in that situation. If the variance is not the same in each group, $\hat{\lambda}_1^{pd}$ should be favored.

4.4 Confidence Intervals

Finally let us investigate confidence intervals for ξ and $\frac{1}{k-1}\sum_{j=2}^{k}\xi_{1j}$ based on $\hat{\lambda}_1^{block}$ and $\hat{\lambda}_1^{pd}$. The usual simple idea is based on the asymptotic normality and estimators for the variances to construct $(1-\alpha)$-confidence intervals, which leads to

$$[\hat{\lambda}_1^{pd} - z_{1-\alpha/2}\sqrt{\hat{\sigma}^2_{\hat{\lambda}_1^{pd}}/n}, \hat{\lambda}_1^{pd} + z_{1-\alpha/2}\sqrt{\hat{\sigma}^2_{\hat{\lambda}_1^{pd}}/n}]$$

and

$$[\hat{\lambda}_1^{block} - z_{1-\alpha/2}\sqrt{\hat{\sigma}^2_{\hat{\lambda}_1^{block}}/n}, \hat{\lambda}_1^{block} + z_{1-\alpha/2}\sqrt{\hat{\sigma}^2_{\hat{\lambda}_1^{block}}/n}]$$

with the variance estimators

$$\hat{\sigma}^2_{\hat{\lambda}_1^{block}} = \frac{\hat{v}^2}{n(k-1)\hat{\bar{g}}^2_{\cdot}(\xi)},$$

where

$$\hat{v}^2 = \frac{1}{4} - \frac{1}{k-1}\sum_{j=2}^{k}(\widehat{G_j(\xi)} - 1/2)^2 - \frac{2}{k-1}\sum_{j=2}^{k-1}\sum_{h=1}^{k-j}(\widehat{G_{j,j+h}(\xi)}\widehat{-G_j(\xi)}\widehat{G_{j+h}(\xi)}),$$

with

$$\widehat{G_j(\xi)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(\tilde{D}_{i:j} \le \hat{\lambda}^{block}),\ j = 2, \ldots, k,$$

and

$$\widehat{G_{j,j+h}(\xi)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(\{\tilde{D}_{i:j} \le \hat{\lambda}^{block}\} \cap \{\tilde{D}_{i:j+h} \le \hat{\lambda}^{block}\}),\ j = 2, \ldots, k-1, h = 1, \ldots, j,$$

and

$$\hat{\sigma}^2_{\hat{\lambda}_1^{pd}} = \frac{1}{(k-1)^2}\sum_{j=2}^{k}\hat{\sigma}^2_{\hat{\lambda}_{1j}} + \frac{2}{(k-1)^2}\sum_{j=2}^{k-1}\sum_{\ell=j+1}^{k}\hat{\gamma}_{\hat{\lambda}_{1j},\hat{\lambda}_{1\ell}},$$

where

$$\hat{\sigma}^2_{\hat{\lambda}_{1j}} = \sqrt{2}\left\{\frac{1}{2n}\sum_{i=1}^{n}[1-\hat{F}_j(X_{i1}-\hat{\lambda}_{1j})]^2 + \frac{1}{2n}\sum_{i=1}^{n}\hat{F}_1^2(X_{ij}+\hat{\lambda}_{1j}) - \frac{1}{n-1}\sum_{i=1}^{n}[\hat{F}_1(X_{i1})\hat{F}_j(X_{ij}) - n\bar{F}_1(X_1)\bar{F}_j(X_j)] - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{F}_1(X_{ij}+\hat{\lambda}_{1j})\right)^2\right\} \Big/ \hat{\varphi}_{1j},\ j=2,\ldots,k,$$

and

$$\hat{\gamma}_{\hat{\lambda}_{1j},\hat{\lambda}_{1\ell}} = \frac{1}{\sqrt{2}}\left\{\frac{1}{n-1}\sum_{i=1}^{n}[(1-\hat{F}_j(X_{i1}))(1-\hat{F}_\ell(X_{i1})) - n(1-\bar{F}_j(X_1))(1-\bar{F}_\ell(X_1))] + \frac{1}{n-1}\sum_{i=1}^{n}[(1-\hat{F}_j(X_{i1}))\hat{F}_1(X_{i\ell}) - n(1-\bar{F}_j(X_1))\bar{F}_1(X_\ell)] + \frac{1}{n-1}\sum_{i=1}^{n}[\hat{F}_1(X_{ij})(1-\hat{F}_\ell(X_{i1})) - n\bar{F}_1(X_j)(1-\bar{F}_\ell(X_1))] + \frac{1}{n-1}\sum_{i=1}^{n}[\hat{F}_1(X_{ij})\hat{F}_1(X_{i\ell})) - n\bar{F}_1(X_j)\bar{F}_1(X_\ell)]\right\} \Big/ \sqrt{\hat{\varphi}_{1j}\hat{\varphi}_{1\ell}},\ j=2,\ldots,k-1,\ell=j+1,\ldots,k,$$

with $\bar{F}_j(X_\ell) = \frac{1}{n}\sum_{i=1}^{n}\hat{F}_j(X_{i\ell}), j,\ell = 1,\ldots,k$. Kernel density estimators with Gaussian kernels and bandwidth selection with the method suggested by Sheather and Jones (1991) were used for $\hat{\varphi}_{1j}, j = 2,\ldots,k$, the estimator for $\int f_1(x+\xi_{1j})f_j(x)\,\mathrm{d}x$, and $\hat{\bar{g}}_{\cdot}(\xi)$.
However simulation results showed that the non-coverage probabilities of these 95 %-confidence intervals deviate from 5 % as the variance estimators are not good estimators for finite sample variances. See Hayoz (2012) for the simulation results and for more details on the variance estimators.

Better confidence intervals can be derived by bootstrap methods. With the R package boot five types of bootstrap confidence intervals by Davison and Hinkley (1997) were simulated; here, we present the three methods that performed best in our simulations. Let $\hat{\lambda}$ be an estimator for an unknown parameter λ. After generating B bootstrap samples we get $\hat{\lambda}_1^*,\ldots,\hat{\lambda}_B^*$. Assume that the mean and variance of the distribution of $\hat{\lambda}$ are $\lambda+\beta$ and σ^2, where β is the bias of $\hat{\lambda}$. We can estimate β by

$\hat{\beta}_{boot} = \frac{1}{B}\sum_{b=1}^{B}\hat{\lambda}_b^* - \hat{\lambda} = \bar{\hat{\lambda}}_{boot} - \hat{\lambda}$ and σ^2 by $\hat{\sigma}^2_{boot} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\lambda}_b^* - \bar{\hat{\lambda}}_{boot})^2$. If we assume that the distribution of $\hat{\lambda}$ is approximately normal, an $1-\alpha$ confidence interval for λ, referred to as *normal bootstrap confidence interval*, is given by

$$[\hat{\lambda} - \hat{\beta}_{boot} - z_{1-\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2_{boot}}, \hat{\lambda} - \hat{\beta}_{boot} + z_{1-\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2_{boot}}].$$

Another method to construct a confidence interval is to use the studentized estimator $Z = \frac{\hat{\lambda}^* - \hat{\lambda}}{\sqrt{\hat{\sigma}^2}}$, where $\hat{\sigma}^2$ is a variance estimator. With the help of B bootstrap copies of $\hat{\lambda}$ and $\hat{\sigma}^2$ we construct $z_1^* = \frac{\hat{\lambda}_1^* - \hat{\lambda}}{\sqrt{\hat{\sigma^2}_1^*}}, \ldots, z_B^* = \frac{\hat{\lambda}_B^* - \hat{\lambda}}{\sqrt{\hat{\sigma^2}_B^*}}$ and use the order statistics $z^*_{(1)} < \ldots < z^*_{(B)}$ to estimate quantiles of the distribution of Z. An $1-\alpha$ confidence interval for λ, referred to as *studentized bootstrap confidence interval*, is now given by

$$[\hat{\lambda} - z^*_{((1-\frac{\alpha}{2})(B+1))}\sqrt{\hat{\sigma}^2}, \hat{\lambda} - z^*_{(\frac{\alpha}{2}(B+1))}\sqrt{\hat{\sigma}^2}].$$

The *basic bootstrap confidence interval* approach is simpler:

$$[2\hat{\lambda} - \hat{\lambda}^*_{((1-\frac{\alpha}{2})(B+1))}, 2\hat{\lambda} - \hat{\lambda}^*_{(\frac{\alpha}{2}(B+1))}].$$

Simulation results of the non-coverage probabilities of the bootstrap confidence intervals are shown in Figs. 4.5 and 4.6 for $\alpha = 0.05$ with $k = 4$ groups, $n = 10$ blocks, $n_{\text{sim}} = 5'000$ and $B = 200$. To check the validity of the results, one scenario (with the normal distribution and the compound symmetry with $\rho = 0.5$) was repeated with B=5000. The results on the non-coverage probability were similar to the results with B=200; differences were noticed in the third digit.

For $\hat{\lambda}_1^{block}$ the non-coverage probabilities of the studentized bootstrap confidence interval are closest to 0.05 if the marginal distributions are symmetric. The non-coverage probabilities of the other bootstrap confidence intervals are much higher than 0.05. If the marginal distributions are asymmetric, the non-coverage probabilities of the normal and studentized bootstrap confidence interval are slightly smaller than 0.05, the non-coverage probabilities of the basic bootstrap confidence interval is higher than 0.05.

For $\hat{\lambda}_1^{pd}$ the non-coverage probabilities of the normal and basic bootstrap confidence interval are closest to 0.05, but the deviation is sometimes big. The non-coverage probabilities of the studentized bootstrap confidence interval are in most situations a bit too low.

In conclusion, for $\hat{\lambda}_1^{block}$ the studentized bootstrap confidence interval is the best method for symmetric marginal distributions, the normal bootstrap confidence interval for asymmetric marginal distributions. For $\hat{\lambda}_1^{pd}$ the normal and basic bootstrap confidence intervals are the best methods.

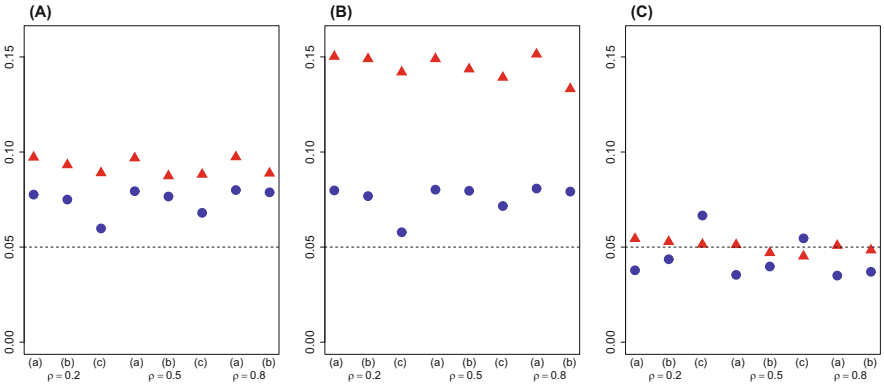

Fig. 4.5 Non-coverage probabilities of normal (**A**), basic (**B**), and studentized (**C**) bootstrap confidence intervals for the estimators $\hat{\lambda}_1^{block}$ (*red filled triangle*) and $\hat{\lambda}_1^{pd}$ (*blue filled circle*) with $F_1 = \ldots = F_k$, an underlying $\mathcal{N}(0, 1)$ distribution, $\alpha = 0.05$, $k = 4, n = 10$ and underlying CS (*a*), AR(1) (*b*), and MA(1) (*c*) covariance structures

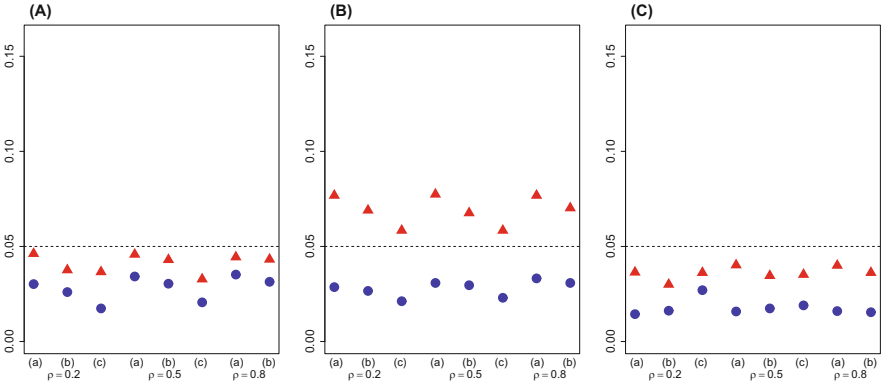

Fig. 4.6 Non-coverage probabilities of normal (**A**), basic (**B**), and studentized (**C**) bootstrap confidence intervals for the estimators $\hat{\lambda}_1^{block}$ (*red filled triangle*) and $\hat{\lambda}_1^{pd}$ (*blue filled circle*) with $F_1 = \ldots = F_k$, an underlying $\log\mathcal{N}(0, 1)$ distribution, $\alpha = 0.05$, $k = 4, n = 10$ and underlying CS (*a*), AR(1) (*b*) and MA(1) (*c*) covariance structures

4.5 Conclusions

We discussed two estimators for the location difference between one group and all the others that are asymptotically normal distributed, consistent, translation invariant, and scale preserving. Both estimators are unbiased and median unbiased for symmetric marginal distributions under the null hypothesis; but these properties do not hold in general. $\hat{\lambda}_1^{block}$ is median unbiased under the null hypothesis or in multiple matching trials and has the smallest variance and MSE, but $\hat{\lambda}_1^{pd}$ is the

better choice if not all groups have the same location. $\hat{\lambda}_1^{pd}$ is unbiased for symmetric marginal distributions or if the groups are exchangeable. The use of confidence intervals for the estimator should be based on bootstrap confidence intervals. For $\hat{\lambda}_1^{block}$ the studentized bootstrap confidence interval is the best method for symmetric marginal distributions, the normal bootstrap confidence interval for asymmetric marginal distributions. For $\hat{\lambda}_1^{pd}$ the normal and basic bootstrap confidence intervals are the best methods.

References

Brunner, E., Domhof, S., Langer, F.: Nonparametric Analysis of Longitudinal Data in Factorial Experiments. Wiley Series in Probability and Statistics. Wiley, New York (2002)

Davis, C.S.: Statistical Methods for the Analysis of Repeated Measurements. Springer, New York (1952).

Davison, A.C., Hinkley, D.V.: Bootstrap Methods and their Application. Cambridge University Press, Cambridge (1997)

Forster, O.: Analysis 2, 5th edn. Friedr. Vieweg und Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden (1999)

Friedman, M.: A comparison of alternative tests of significance for the problem of m rankings. Ann. Math. Stat. **11**, 86–92 (1940)

Hayoz, S.: Verhalten von nichtparametrischen Schätzern und Tests in longitudinalen Versuchen. Master's thesis, University of Bern (2006)

Hayoz, S.: Nonparametric location estimators in the randomized complete block design. Ph.D. thesis, University of Bern (2012)

Hettmansperger, T.P., Sheather, S.J.: Confidence intervals based on interpolated order statistics. Stat. Probabil. Lett. **4**, 75–79 (1986).

Hodges, J.L., Lehmann, E.L.: Estimates of location based on rank tests. Ann. Math. Stat. **34**, 598–611 (1963)

Hollander, M., Pledger, G., Lin, P.: Robustness of the Wilcoxon test to a certain dependency between samples. Ann. Stat. **2**, 177–181 (1974)

Konietschke, F., Bathke, A.C., Hothorn, L.A., Brunner, E.: Testing and estimation of purely nonparametric effects in repeated mesures designs. Comput. Stat. Data Anal. **54**, 1895–1905 (2010)

Koul, H.L.: Weighted Empirical Processes in Dynamic Nonlinear Models, 2nd edn. Lecture Notes in Statistics. Springer, New York (2002)

Lehmann, E.L.: Robust estimation in analysis of variance. Ann. Math. Stat. **34**, 957–966 (1963)

Lehmann, E.L.: Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco, CA (1975)

Sen, P.K.: Asymptotic normality of sample quantiles for *m*-dependent processes. Ann. Math. Stat. **39**, 1724–1730 (1968)

Sheather, S.J., Jones, M.C.: A reliable data-based bandwidth selection method for kernel density estimation. J. R. Stat. Soc. Ser. B **53**, 683–690 (1991)

Yan, J.: Enjoy the joy of copulas: With package copula. J. Stat. Softw. **21**(4), 1–21 (2007)

Chapter 5
Permutation Tests in Linear Regression

Jukka Nyblom

Abstract Exact permutation tests are available only in rather simple linear models. The problem is that, although standard assumptions allow permuting the errors of the model, we cannot permute them in practice, because they are unobservable. Nevertheless, the residuals of the model can be permuted. A proof is given here which shows that it is possible to approximate the unobservable permutation distribution where the true errors are permuted by permuting the residuals. It is shown that approximation holds asymptotically and almost surely for certain quadratic statistics as well as for statistics which are expressible as the maximum of appropriate linear functions. The result is applied to testing the significance of predictors as well as to diagnostic checking of heteroscedasticity, autocorrelation, change-points, and changing regression function. Also a simulation experiment is made in order to evaluate the performance of the proposed tests.

Keywords Autocorrelation • Change-point problem • Control variable • Heteroscedasticity • Simulation

5.1 Introduction

The idea of calculating the significance level of a statistical test by appropriately permuting the observations at hand originates from the writings of Fisher (1935) and Pitman (1937a,b, 1938). During the last two decades we have seen the revival of these ideas, e.g. see Edgington (1995), Manly (1997), Pesarin (2001), and Good (2005) and their references. This is in line with the popularity of other computer intensive methods of which the bootstrap of Efron (1979) is a well-known example.

Sometimes a distinction is made between permutation tests and randomization tests (Kempthorne 1986, p. 524). Randomization tests are related to randomization experiments and the randomization distribution is induced by the true randomization that the experimenter has actually made. This gives a valid method of inference

J. Nyblom (✉)
Department of Mathematics and Statistics, University of Jyväskylä, P.O. Box 35, FIN-40014 Jyväskylä, Finland
e-mail: jukka.nyblom@jyu.fi

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_5

without any assumption on experimental units being a random sample. The permutation tests, however, are confined to random samples. This article is focused on permutation tests in this latter sense. For the basic theory of randomization and permutation tests, see Lehmann and Romano (2005, pp. 176–211).

There are several suggestions in the literature on the proper way of performing the permutation test in regression. Kennedy (1995) gives references and critically discusses several proposals. Different proposals and applications are provided, e.g. by Schmoyer (1994), Kim et al. (2000), Anderson and Robinson (2001), Huh and Jhun (2001), and Anderson and ter Braak (2003). LePage and Podgórski (1996) give some results without assuming a finite variance for errors.

Oja (1987) gives an interesting example of an exact permutation test for treatment effects when (1) the treatment levels are assigned randomly to experimental units, and (2) the values of other predictors are observed before the application of the treatments. The significance may be assessed by comparing the estimated effects with those obtained by rerandomization of the levels. This procedure is more like the randomization tests discussed above, and it is not legitimate in observational studies that are of main interest in this article. Nevertheless, this gives a rationale for the randomization method 1 of Kennedy (1995, p. 90) in some cases.

In this article the problem of a proper permutation test in linear regression is solved as follows. All test statistics are assumed pivotal (independent of the parameters under the null hypothesis). This means that we can write them in terms of the true errors. Hence the observed value of a test statistic is one of the equally likely theoretical values generated by error permutations. But since the errors are unobservable, we have to use the residuals from the model. This amounts, in fact, to using the permuted residuals as new dependent variables that are used to compute the permutation distribution. The method is equivalent to that proposed by Freedman and Lane (1983), see also Kennedy (1995, methods 4 and 5).

The main results show that if the test statistic is a quadratic form or a maximum of a certain linear function, the resulting test approximates the exact permutation test we would use if the errors were observable. The approximation is asymptotic, i.e. the number of cases in regression is supposed to be large compared to the number of predictors, and almost sure in the sense that the approximation holds conditionally on the unobserved errors with probability one. The result generalizes Theorem 1 of Schmoyer (1994). It is also in concordance with the finding of Anderson and Robinson (2001) who show that asymptotically, in a certain special case, the correlation between those two is equal to 1. It is also interesting that we need not assume that the limiting distribution exists neither in the sense of a sampling distribution nor of a permutation distribution.

The article proceeds as follows. First, a brief discussion on exact permutation tests is presented in Sect. 5.2. Then the approximate permutation tests with theorems proving their validity are introduced in Sect. 5.3. The applications are given in Sect. 5.4 including moderate simulation experiments to show how the method works in practice. The examples include the test for a regression coefficient and diagnostic checks (heteroscedasticity, autocorrelation, and changes in regression function). Section 5.5 gives the proofs.

5.2 Exact Permutation Tests

Let $\boldsymbol{y}' = (y_1, \ldots, y_n)$ be a vector of random variables. Assume that under the null hypothesis they are exchangeable, i.e. each permutation $(y_{i_1}, \ldots, y_{i_n})$ follows the same distribution. The alternative hypothesis expresses our beliefs on what kind of deviations from the exchangeability assumption may occur in the data. Let $T(\boldsymbol{y})$ be a suitable test statistic against the alternative we have in mind, the large values indicating against the null hypothesis. Then we can proceed in a distribution-free manner as follows. Compute all $n!$ values of $T(\boldsymbol{y})$ by permuting the coordinates of $\boldsymbol{y}$. Assume that there are $N \leq n!$ distinct values ordered as $T_1 < T_2 < \cdots < T_N$. Assume further that the proportion of all permutations yielding the value T_k is p_k. Fix α and solve $k = k_\alpha$ such that

$$\sum_{i=k}^{N} p_i \leq \alpha \quad \text{and} \quad \sum_{i=k-1}^{N} p_i > \alpha.$$

Then the exact level α test is obtained by the following decision rule: reject the null hypothesis with probability 1 if $T(\boldsymbol{y}) \geq T_k$, and with probability $\alpha - \sum_{i=k}^{N} p_i$ if $T(\boldsymbol{y}) = T_{k-1}$. One seldom uses this randomized procedure in practice but only computes the so-called p value that equals the proportion of those values of T_k that are larger than or equal to $T(\boldsymbol{y})$. As an example consider a simple linear regression. The exact p value for the least squares estimate of the slope can be based on the distribution of all least squares estimates derived from the permuted dependent variable.

The value $n!$ may well be too large for a complete enumeration. Then we can compute an approximate p value by a Monte Carlo method. The resulting Monte Carlo test is not valid only as an approximation to the exact permutation test but also serves as a test of its own right. This is seen in the next theorem the proof of which is a modification of that given by Schmoyer (1994, Sect. 2).

Theorem 5.1 *Let Σ be a random sample (with replacement) of size m from the population of all permutations. If the random variables $y_1, \ldots, y_n$ are exchangeable, then*

$$P\left(\frac{1}{m}\sum_{\pi \in \Sigma} I[T(y) \leq T(y_\pi)] \leq \alpha\right) \leq \frac{[m\alpha] + 1}{m + 1},$$

where $I[\cdot]$ is an indicator function, $\boldsymbol{y}_\pi$ denotes the permuted vector obtained from $\boldsymbol{y}$ and $[x]$ denotes the largest integer not greater than x.

Proof By the exchangeability of the coordinates of $\boldsymbol{y}$ the set $\{T(\boldsymbol{y}), T(\boldsymbol{y}_\pi), \pi \in \Sigma\}$ consists of independent and identically distributed random variables. The expression

$$1 + \sum I[T(\boldsymbol{y}) \leq T(\boldsymbol{y}_\pi)]$$

gives the reversed rank (rank 1 is given to the largest value) of $T(\boldsymbol{y})$ in the set. In case of ties this procedure gives the largest reversed rank for each tied observation. If ties are broken randomly, then the probability of the event that the rank is at most r is exactly $r/(m+1)$. Thus the probability that the largest rank is at most r cannot exceed $r/(m+1)$. This proves the inequality. □

5.3 Approximate Permutation Tests

Let us define the general linear model

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{5.1}$$

where $\mathbf{1}$ is a column of ones, and $\boldsymbol{X}$ is an $n \times q$ matrix of predictors with $[\mathbf{1}, \boldsymbol{X}]$ being of full rank. The scalar β_0 is the intercept and $\boldsymbol{\beta}$ the vector of regression coefficients. Additionally, in order to alleviate the asymptotic analysis, we assume that the columns of $\boldsymbol{X}$ are centered. The rows of $\boldsymbol{X}$ are denoted by $\boldsymbol{x}_i'$ (a row vector). The most common assumption for the errors $(e_1, \ldots, e_n) = \boldsymbol{e}'$ is that they are independent and identically distributed with mean zero and variance σ^2. However, independence is replaced by exchangeability, because the proofs of Theorems 5.2 and 5.3 remain exactly the same. We proceed under this assumption. Let us further denote the projection matrix $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ by $\boldsymbol{H}$.

In applications we may be interested in testing for the significance of some additional predictor variables, or in making diagnostic checks (e.g., heteroscedasticity, autocorrelation, changing regression coefficients, etc.). Whatever the test may be we assume that it is free of the coefficients β_0 and $\boldsymbol{\beta}$. This achieved if the test is based on the least squares residuals $\hat{\boldsymbol{e}} = \boldsymbol{y} - \bar{y}\mathbf{1} - \boldsymbol{H}\boldsymbol{y} = \boldsymbol{Q}\boldsymbol{y}$ with $\boldsymbol{Q} = \boldsymbol{I} - n^{-1}\mathbf{1}\mathbf{1}' - \boldsymbol{H}$. Under (5.1) we can also write that $\hat{\boldsymbol{e}} = \boldsymbol{Q}\boldsymbol{e} = \boldsymbol{e} - \bar{e}\mathbf{1} - \boldsymbol{H}\boldsymbol{e}$. Sometimes it is desirable to stress the dependence on n and write $\boldsymbol{H}_n, \boldsymbol{Q}_n, \ldots$ instead of $\boldsymbol{H}, \boldsymbol{Q}, \ldots$.

Let $T(\boldsymbol{Q}\boldsymbol{y}) = T(\hat{\boldsymbol{e}})$ be a test statistic for (5.1) against some alternative. Under the null hypothesis we can write $T(\boldsymbol{Q}\boldsymbol{y}) = T(\boldsymbol{Q}\boldsymbol{e})$. Suppose for a moment that the true errors were observable, then we should permute the coordinates of $\boldsymbol{e}$ in $T(\boldsymbol{Q}\boldsymbol{e})$. If these permuted vectors are denoted by $\boldsymbol{e}_\pi$ with π running over the set Π of all permutations, the reference distribution is obtained through the values

$$T(\boldsymbol{Q}\boldsymbol{e}_\pi), \quad \pi \in \Pi$$

with the assignment of the probability $1/n!$ for each permutation. By this procedure we would get an exact permutation test. But we cannot observe $\boldsymbol{e}$. Therefore we try to approximate it by means of

$$T(\boldsymbol{Q}\hat{\boldsymbol{e}}_\pi), \quad \pi \in \Pi.$$

The procedure means that we first find the residuals, permute them, and then use them as new dependent variables in the null hypothesis regression. We could equally well use the pseudo values $\hat{\boldsymbol{y}}_\pi = \hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}_\pi$ with $\hat{\boldsymbol{y}} = \bar{y}\mathbf{1} + \boldsymbol{H}\boldsymbol{y}$, as Freedman and Lane (1983) do, but adding the fitted value $\hat{\boldsymbol{y}}$ is not necessary as $\boldsymbol{Q}\hat{\boldsymbol{y}} = \mathbf{0}$. Schmoyer (1994) suggests using $T(\hat{\boldsymbol{e}}_\pi)$.

Let us now consider the important special cases of quadratic forms that are of particular interest in linear models. Let us assume that the test statistic is of the form

$$T(\hat{\boldsymbol{e}}) = \frac{\hat{\boldsymbol{e}}'\boldsymbol{A}\hat{\boldsymbol{e}}}{\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}}.$$

When using the corresponding permutation test, we take the numerator as our test statistic, because the denominator is merely for scaling. Thus, the test statistic is

$$T(\boldsymbol{Q}\hat{\boldsymbol{e}}_\pi) = \hat{\boldsymbol{e}}_\pi'\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}\hat{\boldsymbol{e}}_\pi. \tag{5.2}$$

Let us find its expectation in two steps, first over the permutation distribution and then over the sampling distribution. The covariance matrix for $\hat{\boldsymbol{e}}_\pi$ over the permutation distribution is after some algebra as follows:

$$\hat{\sigma}^2\boldsymbol{V} = \mathrm{cov}_\pi(\hat{\boldsymbol{e}}_\pi) = \hat{\sigma}^2\frac{n}{n-1}\left[\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right], \quad \hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\hat{e}_i^2. \tag{5.3}$$

Then, by $\boldsymbol{Q}\mathbf{1} = \mathbf{0}$, we have

$$E\left[\hat{\boldsymbol{e}}_\pi'\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}\hat{\boldsymbol{e}}_\pi\right] = E(\hat{\sigma}^2)\frac{n}{n-1}\mathrm{tr}(\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}) = \sigma^2\frac{n-q-1}{n-1}\mathrm{tr}(\boldsymbol{A}\boldsymbol{Q}).$$

As a comparison, under the sampling distribution we have

$$E\left[\hat{\boldsymbol{e}}'\boldsymbol{A}\hat{\boldsymbol{e}}\right] = \sigma^2\mathrm{tr}(\boldsymbol{A}\boldsymbol{Q}).$$

If $\mathrm{tr}(\boldsymbol{A}\boldsymbol{Q}) \neq 0$, the bias correction can be done by multiplying the quadratic permutation statistic with the factor

$$\frac{n-1}{n-q-1}. \tag{5.4}$$

A fairly general theorem is now given showing the asymptotic validity of the proposed permutation tests. The proof of the theorem and the necessary lemmas are postponed to Sect. 5.5 at the end.

Theorem 5.2 *Let $\boldsymbol{B}_n = (b_{ijn})$ be a sequence of symmetric $n \times n$ matrices such that $\boldsymbol{B}_n \mathbf{1}_n = \mathbf{0}$, for all n, and*

$$\sup_n \operatorname{tr}(\boldsymbol{B}_n^2) = \sup_n \sum_{i=1}^n \sum_{j=1}^n b_{ijn}^2 \leq C.$$

Assume that (5.1) holds with exchangeable errors and with $Ee_1^6 < \infty$, then for all $\varepsilon > 0$

$$\lim_{n\to\infty} P_\pi(|\hat{\boldsymbol{e}}'_\pi \boldsymbol{B}_n \hat{\boldsymbol{e}}_\pi - \boldsymbol{e}'_\pi \boldsymbol{B}_n \boldsymbol{e}_\pi| > \varepsilon) = 0 \quad a.s.,$$

where P_π denotes the conditional probability induced by the uniform permutation distribution, and "a.s." qualifies the almost sure convergence with respect to the infinite sequence of true errors $e_1, e_2, \ldots$.

Remark 5.1 The assumption of the finite sixth moment is purely technical. Its purpose is to give a simple proof for Lemma 5.1 in Sect. 5.5.

The next theorem is useful in change-point problems in regression. The approach based on sampling distribution under normal errors is difficult even in the case of linear time trend regression (Kim and Siegmund 1989). The permutation approach is, nevertheless, generally applicable.

Theorem 5.3 *Let $z_{1n}, \ldots, z_{nn}$ with $n > q + 1$ be a triangular sequence of numbers such that $\sum_{j=1}^n z_{jn}^2 = 1$. Denote $\boldsymbol{z}'_{kn} = (z_{1n}, \ldots, z_{kn}, 0, \ldots, 0)$ and assume that there are integers $m_0(n)$ and $m_1(n)$ depending on n such that for some positive constants M and b we have*

$$\inf_{n>M} \min_{m_0(n)\leq k\leq m_1(n)} \boldsymbol{z}'_{kn} \boldsymbol{Q}_n \boldsymbol{z}_{kn} \geq b. \tag{5.5}$$

Assume that (5.1) holds with exchangeable errors and with $Ee_1^6 < \infty$, then for all $\varepsilon > 0$

$$\lim_{n\to\infty} P_\pi \left(\max_{m_0(n)\leq k\leq m_1(n)} \frac{|\boldsymbol{z}'_{kn} \boldsymbol{Q}_n (\hat{\boldsymbol{e}}_\pi - \boldsymbol{e}_\pi)|}{\sqrt{\boldsymbol{z}'_{kn} \boldsymbol{Q}_n \boldsymbol{z}_{kn}}} > \varepsilon \right) = 0, \quad \text{a.s.}$$

Remark 5.2 Let X_n and $\hat{X}_n$ be the permutation statistics involving true errors and the residuals, respectively. Let the assumptions of either Theorem 5.2 or Theorem 5.3 hold. If X_n has a continuous limiting distribution, then also $\hat{X}_n$ has the same limiting distribution. The continuity of the common limiting distribution implies that both convergences are uniform (Rao 1973, p. 120), and therefore

$$\lim_{n\to\infty} \sup_x |P_\pi(X_n \leq x) - P_\pi(\hat{X}_n \leq x)| = 0 \quad \text{a.s.}$$

5.4 Applications

We explore the performance of permutation tests with the help of real data and simulation experiments. The real data problems focus on diagnostic checks concerning heteroscedasticity, autocorrelation, and change-points. Additionally, tests for linear hypothesis and for constancy of linear trend are treated with artificial data. In simulation experiments the response variables are generated from various distributions: normal, Laplace, t-distribution with 4 degrees of freedom (though it does not satisfy the moment condition of Theorems 2 and 3) and log-normal distribution. The first three are all symmetric with varying tail behavior. The chosen log-normal distribution is moderately skew: after log-transformation the distribution is normal with zero mean and variance 0.09.

In each application the simulations consist of 10,000 replicates of permutation tests, and each test is based on 999 random permutations. The p-values are then estimated with the help of control variables. The natural control variables are the p values of the exact permutation test that is known to be uniformly distributed. The p values from the test under evaluation, and the p values from the corresponding exact test are cross-classified into a 4×4 table. The row marginals give the frequencies of the exact permutation tests classified according to the intervals $[0, 0.01]$, $(0.01, 0.05]$, $(0.05, 0.10]$, $(0.10, 1]$. The column marginals give the frequencies for the test under evaluation under the same classification. Denote the cell frequencies and the corresponding cell probabilities as f_{ij}, θ_{ij}, respectively, $i, j = 1, 2, 3, 4$. For example, f_{23} is the number of those replications, where the exact test yields the p value in $(0.01, 0.05]$, and the p value of the test under evaluation is in $(0.05, 0.10]$, both occurring in the same replicate. Because the p values of the exact test are uniformly distributed, the row marginals $f_{i\cdot}$, $i = 1, 2, 3, 4$ related to the exact test are known to be multinomially distributed with the known probabilities $\alpha_1 = 0.01, \alpha_2 = 0.04, \alpha_3 = 0.05, \alpha_4 = 0.90$. Conditionally on the row sum $f_{i\cdot}$, the row frequencies $f_{i1}, f_{i2}, f_{i3}, f_{i4}$ follow the multinomial distribution $\mathrm{MN}(f_{i\cdot}; \theta_{i1}/\alpha_i, \theta_{i2}/\alpha_i, \theta_{i3}/\alpha_i, \theta_{i4}/\alpha_i)$. Therefore, after some algebra using the restrictions $\sum_j \theta_{ij} = \alpha_i$, $i = 1, 2, 3, 4$, we have conditionally unbiased estimates for the cell probabilities

$$\hat{\theta}_{ij} = \alpha_i \frac{f_{ij}}{f_{i\cdot}}, \quad i, j = 1, 2, 3, 4.$$

Let $\sum_{ij} \hat{\theta}_{ij} = \hat{\theta}_{\cdot j}$. Then the size estimates, at the nominal 1 %, 5 %, and 10 % levels are $\hat{\theta}_{\cdot 1}$, $\hat{\theta}_{\cdot 1} + \hat{\theta}_{\cdot 2}$, $\hat{\theta}_{\cdot 1} + \hat{\theta}_{\cdot 2} + \hat{\theta}_{\cdot 3}$, respectively. The standard errors for them can be

found by noting that conditionally on the row sums, the frequencies at the different rows are independent multinomials. Therefore

$$\text{var}\,(\hat{\theta}_{\cdot 1} \mid f_{1\cdot}, f_{2\cdot}, f_{3\cdot}, f_{4\cdot}) = \sum_{i=1}^{4} \frac{\theta_{i1}(\theta_{i2}+\theta_{i3}+\theta_{i4})}{f_{i\cdot}},$$

$$\text{var}\,(\hat{\theta}_{\cdot 1}+\hat{\theta}_{\cdot 2} \mid f_{1\cdot}, f_{2\cdot}, f_{3\cdot}, f_{4\cdot}) = \sum_{i=1}^{4} \frac{(\theta_{i1}+\theta_{i2})(\theta_{i3}+\theta_{i4})}{f_{i\cdot}},$$

$$\text{var}\,(\hat{\theta}_{\cdot 1}+\hat{\theta}_{\cdot 2}+\hat{\theta}_{\cdot 3} \mid f_{1\cdot}, f_{2\cdot}, f_{3\cdot}, f_{4\cdot}) = \sum_{i=1}^{4} \frac{(\theta_{i1}+\theta_{i2}+\theta_{i3})\theta_{i4}}{f_{i\cdot}}.$$

The standard errors are obtained by inserting the estimates in these formulas and taking the square root.

5.4.1 Additional Predictors

Let $\boldsymbol{Z}$ be an $n \times r$ matrix, of full rank, consisting of the values of the additional predictors, i.e. our model is now $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{e}$ with errors as before. Our null hypothesis is $H_0 : \boldsymbol{\gamma} = \boldsymbol{0}$. The alternative hypothesis is $H_A : \boldsymbol{\gamma} \neq \boldsymbol{0}$. Denote by $SSE_0 = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$ the residual sum of squares under the null hypothesis and by SSE_1 the corresponding sum of squares under the alternative hypothesis. Then the usual F-test rejects with large values of

$$F = \frac{(SSE_0 - SSE_1)/r}{SSE_1/(n-q-r-1)}.$$

After an easy algebra we find that $SSE_1 = SSE_0 - \hat{\boldsymbol{e}}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{QZ})^{-1}\boldsymbol{Z}'\hat{\boldsymbol{e}}$. Hence F-test is equivalent to the test rejecting as

$$\frac{\hat{\boldsymbol{e}}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{QZ})^{-1}\boldsymbol{Z}'\hat{\boldsymbol{e}}}{\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}} > c. \tag{5.6}$$

Under the null hypothesis $\hat{\boldsymbol{e}} = \boldsymbol{Qe}$. As the numerator of (5.6) is sufficient, the permutation test statistic is

$$\hat{\boldsymbol{e}}'_{\pi}\boldsymbol{QZ}(\boldsymbol{Z}'\boldsymbol{QZ})^{-1}\boldsymbol{Z}'\boldsymbol{Q}\hat{\boldsymbol{e}}_{\pi}.$$

The matrix $\boldsymbol{B}$ of Theorem 5.2 specifies here to $\boldsymbol{QZ}(\boldsymbol{Z}'\boldsymbol{QZ})^{-1}\boldsymbol{Z}'\boldsymbol{Q}$. Since the sum of its squared elements equals r, the permutation test is asymptotically valid. The procedure is equivalent to that suggested by Freedman and Lane (1983) and

Table 5.1 Estimated relative rejection frequencies for the interaction test

Distribution	Method	1 %	5 %	10 %
Normal	Uncorrected	0.0149 (0.0153)	0.0592 (0.0623)	0.1095 (0.1123)
	Corrected	0.0061 (0.0062)	0.0507 (0.0536)	0.0982 (0.1011)
	Schmoyer	0.0000 (0.0000)	0.0329 (0.0346)	0.0830 (0.0860)
	Exact perm.	0.0098	0.0535	0.1028
Laplace	Uncorrected	0.0206 (0.0209)	0.0583 (0.0573)	0.1092 (0.1079)
	Corrected	0.0103 (0.0106)	0.0521 (0.0512)	0.1008 (0.0995)
	Schmoyer	0.0000 (0.0000)	0.0350 (0.0348)	0.0866 (0.0854)
	Exact perm.	0.0109	0.0490	0.0987
$t(4)$	Uncorrected	0.0213 (0.0207)	0.0581 (0.0552)	0.1065 (0.1038)
	Corrected	0.0120 (0.0118)	0.0515 (0.0488)	0.0982 (0.0955)
	Schmoyer	0.0000 (0.0000)	0.0329 (0.0315)	0.0853 (0.0825)
	Exact perm.	0.0105	0.0469	0.0973
Log-normal	Uncorrected	0.0252 (0.0252)	0.0554 (0.0570)	0.0956 (0.0978)
	Corrected	0.0167 (0.0166)	0.0515 (0.0529)	0.0905 (0.0927)
	Schmoyer	0.0000 (0.0000)	0.0334 (0.0339)	0.0859 (0.0880)
	Exact perm.	0.0090	0.0516	0.1024

The simulated p values of the exact permutation test is used as the control in estimating the sizes of the other tests (raw rejection frequencies are in the parentheses). Standard errors of the estimates from the control variable method are at most 0.0013

discussed by Kennedy (1995, methods 4 and 5, p. 90). Schmoyer's approach leads to using $\hat{\boldsymbol{e}}_\pi' \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z})^{-1}\boldsymbol{Z}\hat{\boldsymbol{e}}_\pi$.

When the hypotheses are $H_0 : \boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ and $H_A : \boldsymbol{\gamma} \neq \boldsymbol{\gamma}_0$, write $\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}_0 = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) + \boldsymbol{e}$. Plainly, the test can be performed replacing $\boldsymbol{y}$ with $\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}_0$.

The simulation experiment consists of generating artificial data of the 2×2 classification with nine observations in a cell, i.e. $n = 36$. Table 5.1 shows relative rejection frequencies of the true null hypothesis that there is no interaction. The uncorrected permutation test is somewhat oversized and Schmoyer's test undersized, whereas the corrected permutation test performs best.

5.4.2 *Heteroscedasticity*

Cook and Weisberg (1983) have derived a general score test against heteroscedastic errors in regression models. A simple special case is treated here. Assume the regression model (5.1), where the errors are assumed to have $\text{var}(e_i) = \sigma^2 e^{\lambda z_i}$, $i = 1, \ldots, n$. The hypotheses are $H_0 : \lambda = 0$ and $H_A : \lambda > 0$. Under the normal

distribution the score test rejects when

$$S = \frac{\sum (z_i - \bar{z})\hat{e}_i^2}{\hat{\sigma}^2 \sqrt{2 \sum (z_i - \bar{z})^2}} > c. \tag{5.7}$$

Let $\boldsymbol{A}$ be the diagonal matrix of the elements $(z_i - \bar{z})/\sqrt{2\sum (z_j - \bar{z})^2}$, $i = 1, \ldots, n$. Then $\boldsymbol{B} = \boldsymbol{QAQ}$ clearly satisfies the condition of Theorem 5.2.

Cook and Weisberg (1982, p. 66) give a data set of black cherry trees (originally from Ryan et al. 1976). The data consist of three measurements of 31 trees: volume, diameter, and height. We model the cube root of the volume to depend linearly on the height and the diameter, then the test (5.7) with the height variable taking the role of z_i's. The exact p value under normal error assumption is obtained by the Imhof technique (Imhof 1961) and appears to be 0.023. The permutation test of the type (5.2) with 9999 random permutations gives the p values 0.0118 and 0.0064 with and without the correction (5.4), respectively. Schmoyer's permutation test yields the p value 0.0095. All tests reject the null hypothesis.

In the simulation study the values z_i in (5.7) are chosen as the height variable. The results are found in Table 5.2. Again, the corrected permutation test performs better than the uncorrected one. Under the normal distribution the corrected permutation test performs best, but under other distributions Schmoyer's test hits closest to the nominal figures.

Table 5.2 Estimated relative rejection frequencies for the heteroscedasticity test

Distribution	Method	1 %	5 %	10 %
Normal	Uncorrected	0.0136 (0.0140)	0.0635 (0.0644)	0.1176 (0.1180)
	Corrected	0.0082 (0.0084)	0.0494 (0.0502)	0.0992 (0.0997)
	Schmoyer	0.0088 (0.0090)	0.0483 (0.0491)	0.0956 (0.0962)
	Exact perm.	0.0103	0.0511	0.1004
Laplace	Uncorrected	0.0209 (0.0216)	0.0710 (0.0728)	0.1282 (0.1300)
	Corrected	0.0138 (0.0142)	0.0571 (0.0587)	0.1119 (0.1138)
	Schmoyer	0.0095 (0.0098)	0.0482 (0.0497)	0.0955 (0.0974)
	Exact perm.	0.0103	0.0519	0.1019
$t(4)$	Uncorrected	0.0214 (0.0223)	0.0723 (0.0746)	0.1263 (0.1281)
	Corrected	0.0141 (0.0148)	0.0577 (0.0598)	0.1078 (0.1097)
	Schmoyer	0.0105 (0.0110)	0.0476 (0.0495)	0.0947 (0.0967)
	Exact perm.	0.0105	0.0527	0.1018
Log-normal	Uncorrected	0.0205 (0.0222)	0.0724 (0.0739)	0.1259 (0.1277)
	Corrected	0.0138 (0.0152)	0.0562 (0.0577)	0.1082 (0.1100)
	Schmoyer	0.0111 (0.0123)	0.0507 (0.0522)	0.1002 (0.1019)
	Exact perm.	0.0119	0.0512	0.1019

The simulated p values of the exact permutation test is used as the control in estimating the sizes of the other tests (raw rejection frequencies are in the parentheses). Standard errors of the estimates from the control variable method are at most 0.0017

5.4.3 Autocorrelation

Assume the regression model (5.1), but now the errors e_i satisfy

$$e_i = \rho e_{i-1} + u_i, \quad i = 2, \ldots, n,$$

where the disturbances u_i are mutually independent with $E(u_i) = 0$ and $\text{var}(u_i) = \sigma^2$, and e_1 is independent of the u_i's with $E(e_1) = 0$, $\text{var}(e_1) = \sigma^2/(1-\rho^2)$, $|\rho| < 1$. The hypotheses are $H_0 : \rho = 0$ and $H_A : \rho > 0$. Under the normal distribution the score test rejects when the first residual autocorrelation coefficient r_1 satisfies

$$r_1 = \frac{n^{-1} \sum \hat{e}_i \hat{e}_{i+1}}{\hat{\sigma}^2} > c. \tag{5.8}$$

Multiplying r_1 by $\sqrt{n}$ we find that the matrix $\boldsymbol{B} = \boldsymbol{QAQ}$ with $\boldsymbol{A}$ having $1/(2\sqrt{n})$ in the first subdiagonal above and below the main diagonal and zero elsewhere. Thus, $\text{tr}(\boldsymbol{B}^2) \leq \text{tr}(\boldsymbol{A}^2) < 1$.

Koutsoyiannis (1973, p. 221) gives the yearly observations on imports and gross national product (GNP) in the UK 1950–1969. Regressing the imports on GNP, $r_1 = 0.199$. The permutation tests give p values 0.0139 and 0.0094 with and without correction, respectively, with 10,000 random permutations. Schmoyer's test yields $p = 0.0239$. If we assume normality, the Imhof technique is applicable also here. It gives $p = 0.06$. Here all permutation tests clearly reject the null hypothesis, while the conclusion under the normal theory test is less clear.

In simulation experiments the response variable is simulated as in Sect. 5.4.2 while the predictor variable is kept as GNP. Results are in Table 5.3. In this case the corrected permutation test performs best. Schmoyer's test gives rejection frequencies well below the nominal levels.

5.4.4 Change-Point Problem

Assume that the observations in the regression model (5.1) are ordered by time (or by some other principle). We suspect that one regression coefficient changes at time k. Denote by $\boldsymbol{z} = (z_1, \ldots, z_n)'$ the column of $[\boldsymbol{1}, \boldsymbol{X}]$ corresponding the dubious regression coefficient. The hypotheses are now

$$\begin{aligned} H_0 &: \text{the coefficient of } z_i \text{ is } \gamma, \quad i = 1, \ldots, n, \\ H_A &: \text{for some } k \text{ the coefficient of } z_i \text{ is } \gamma, \quad i = 1, \ldots, k, \\ &\quad \text{and } \gamma' \neq \gamma, \quad i = k+1, \ldots, n. \end{aligned}$$

If k were known, we could use the t-statistic

$$|t_k| = \frac{|\boldsymbol{z}_k' \boldsymbol{Q} \boldsymbol{y}|}{s_k \sqrt{\boldsymbol{z}_k' \boldsymbol{Q} \boldsymbol{z}_k}},$$

Table 5.3 Estimated relative rejection frequencies for the autocorrelation test

Distribution	Method	1 %	5 %	10 %
Normal	Uncorrected	0.0132 (0.0128)	0.0560 (0.0578)	0.1062 (0.1074)
	Corrected	0.0093 (0.0089)	0.0497 (0.0514)	0.0989 (0.1002)
	Schmoyer	0.0044 (0.0041)	0.0267 (0.0271)	0.0620 (0.0637)
	Exact perm.	0.0093	0.0519	0.1012
Laplace	Uncorrected	0.0127 (0.0114)	0.0557 (0.0508)	0.1070 (0.1018)
	Corrected	0.0094 (0.0085)	0.0497 (0.0451)	0.0974 (0.0922)
	Schmoyer	0.0037 (0.0033)	0.0266 (0.0240)	0.0607 (0.0557)
	Exact perm.	0.0090	0.0451	0.0947
$t(4)$	Uncorrected	0.0118 (0.0106)	0.0556 (0.0539)	0.1068 (0.1046)
	Corrected	0.0097 (0.0086)	0.0490 (0.0474)	0.0999 (0.0978)
	Schmoyer	0.0048 (0.0042)	0.0265 (0.0251)	0.0612 (0.0595)
	Exact perm.	0.0087	0.0484	0.0978
Log-normal	Uncorrected	0.0126 (0.0118)	0.0582 (0.0606)	0.1073 (0.1105)
	Corrected	0.0095 (0.0087)	0.0503 (0.0523)	0.1003 (0.1034)
	Schmoyer	0.0055 (0.0049)	0.0274 (0.0277)	0.0621 (0.0646)
	Exact perm.	0.0089	0.0523	0.1032

The simulated p values of the exact permutation test are used as the control in estimating the sizes of the other tests (raw rejection frequencies are in the parentheses). Standard errors of the estimates from the control variable method are at most 0.0010

where $\boldsymbol{z}_k$ is the vector obtained from $\boldsymbol{z}$ by replacing the coordinates $z_{k+1}, \ldots, z_n$ by zeros and s_k^2 the unbiased residual variance from the augmented regression with $\boldsymbol{X}$ replaced by $[\boldsymbol{X}, \boldsymbol{z}_k]$ in (5.1). Since the t-test is equivalent to the likelihood ratio test (with given k), the likelihood ratio test against alternatives $m_0 \le k \le m_1$ rejects when

$$\max_{m_0 \le k \le m_1} |t_k| > c.$$

Recall that $\boldsymbol{X}$ is $n \times q$. An easy calculation gives

$$(n-q-2)s_k^2 = (n-q-1)s^2 - (\boldsymbol{z}_k'\boldsymbol{Q}\boldsymbol{y})^2/\boldsymbol{z}_k'\boldsymbol{Q}\boldsymbol{z}_k,$$

where s^2 is the unbiased residual variance under the null hypothesis of no change point. From this we find that the t-test is equivalent to the test where s_k is replaced by s. Thus the likelihood ratio test against change-point on the interval $[m_0, m_1]$ rejects when

$$\max_{m_0 \le k \le m_1} \frac{|\boldsymbol{z}_k'\boldsymbol{Q}\boldsymbol{y}|}{s\sqrt{\boldsymbol{z}_k'\boldsymbol{Q}\boldsymbol{z}_k}} > c, \tag{5.9}$$

Since $\boldsymbol{Q}\boldsymbol{y} = \boldsymbol{Q}\boldsymbol{e}$ under the null hypothesis of no change, we can find the p value by permutation whenever $\boldsymbol{z}$ satisfies the condition (5.5) of Theorem 5.3. The condition

roughly means that we exclude the possibility of the change-point occurring too early or too late and that the meaning of "too early" as well as "too late" depends on the particular explanatory variable. Note also that the value of (5.9) remains the same even if we replace z_k by $z_{kn} = z_k/\sqrt{z'z}$.

Chen (1998) applies Schwarz's information criterion (Schwarz 1978) for detecting the change-point of a simple linear regression model. The monthly dollar volume of sales on the Boston Stock Exchange is the response variable and the combined New York American Stock Exchange is the predictor. The data covers the time span from January of 1967 to November of 1969, 35 cases altogether, and is originally from Holbert (1982). Chen's analysis suggests that from December of 1968 onwards the slope is different.

The corrected and uncorrected permutation tests yield p values 0.095 and 0.083, respectively, with 9999 random permutations. Because the test statistic is based on a linear function, the root of the correction factor (5.4) is used. Schmoyer's test gives p value 0.125. Thus, the discovered change-point is not entirely convincing. The interval of maximization in (5.9) is $(4, 32)$.

In simulation experiments the response is generated from the same distributions as before, and the predictor is kept as in the original data. Table 5.4 shows results. Again, the corrected permutation test performs best. As in the autocorrelation test Schmoyer's test remains substantially below the nominal level.

Table 5.4 Estimated relative rejection frequencies for the change-point test

Distribution	Method	1 %	5 %	10 %
Normal	Uncorrected	0.0115 (0.0111)	0.0569 (0.0542)	0.1106 (0.1062)
	Corrected	0.0093 (0.0090)	0.0501 (0.0477)	0.1004 (0.0962)
	Schmoyer	0.0040 (0.0039)	0.0261 (0.0250)	0.0646 (0.0616)
	Exact perm.	0.0097	0.0476	0.0955
Laplace	Uncorrected	0.0121 (0.0118)	0.0572 (0.0600)	0.1093 (0.1096)
	Corrected	0.0107 (0.0103)	0.0499 (0.0527)	0.1003 (0.1009)
	Schmoyer	0.0052 (0.0050)	0.0305 (0.0319)	0.0678 (0.0700)
	Exact perm.	0.0095	0.0533	0.1003
$t(4)$	Uncorrected	0.0117 (0.0119)	0.0566 (0.0557)	0.1093 (0.1088)
	Corrected	0.0088 (0.0090)	0.0498 (0.0489)	0.1002 (0.0997)
	Schmoyer	0.0052 (0.0053)	0.0303 (0.0300)	0.0678 (0.0670)
	Exact perm.	0.0103	0.0490	0.0995
Log-normal	Uncorrected	0.0123 (0.0127)	0.0580 (0.0586)	0.1109 (0.1105)
	Corrected	0.0093 (0.0097)	0.0505 (0.0512)	0.0998 (0.0995)
	Schmoyer	0.0035 (0.0036)	0.0280 (0.0286)	0.0659 (0.0663)
	Exact perm.	0.0104	0.0508	0.0996

The simulated p values of the exact permutation test is used as the control in estimating the sizes of the other tests (raw rejection frequencies are in the parentheses). Standard errors of the estimates from the control variable method are at most 0.0011

5.4.5 Constancy of Linear Trend

Consider the local linear trend model

$$y_i = \mu_i + e_i$$
$$\mu_i = \mu_{i-1} + \beta + d_i, \quad i = 1, \ldots, n,$$

where the errors e_i and d_i are all mutually independent with mean zero and variances σ^2 and σ_d^2, respectively. Plainly, the trend is constant when $\sigma_d^2 = 0$. Then the model takes the form

$$y_i = \mu_0 + \beta i + e_i.$$

The test proposed by Nyblom (1986) rejects the constant trend hypothesis, when

$$L = \frac{\sum_{k=1}^{n} \left(\sum_{i=1}^{k} \hat{e}_i \right)^2}{n^2 \hat{\sigma}^2} > c.$$

The test statistic L has a nonstandard limiting distribution. The permutation tests are compared through artificial data, with $n = 35$, using the same distributions as before. The results are in Table 5.5. The corrected permutation performs best again. Schmoyer's test yields no rejections, and the figures are therefore omitted. The poor performance of Schmoyer's test pertains to the fact that the permutation distribution of L depends on whether $\sum t\hat{e}_{\pi t} = 0$ holds or not. This has counterparts in sampling

Table 5.5 Estimated relative rejection frequencies for the changing trend test

Distribution	Method	1 %	5 %	10 %
Normal	Uncorrected	0.0113 (0.0112)	0.0553 (0.0540)	0.1064 (0.1034)
	Corrected	0.0093 (0.0092)	0.0496 (0.0485)	0.0988 (0.0959)
	Exact perm.	0.0099	0.0489	0.0969
Laplace	Uncorrected	0.0105 (0.0099)	0.0551 (0.0525)	0.1081 (0.1056)
	Corrected	0.0090 (0.0085)	0.0501 (0.0476)	0.0992 (0.0967)
	Exact perm.	0.0094	0.0474	0.0975
$t(4)$	Uncorrected	0.0126 (0.0118)	0.0566 (0.0538)	0.1094 (0.1067)
	Corrected	0.0099 (0.0092)	0.0495 (0.0468)	0.0982 (0.0955)
	Exact perm.	0.0093	0.0471	0.0973
Log-normal	Uncorrected	0.0111 (0.0102)	0.0571 (0.0537)	0.1089 (0.1077)
	Corrected	0.0093 (0.0086)	0.0509 (0.0475)	0.0994 (0.0979)
	Exact perm.	0.0092	0.0462	0.0988

The simulated p values of the exact permutation test are used as the control in estimating the sizes of the other tests (raw rejection frequencies are in the parentheses). Standard errors of the estimates from the control variable method are at most 0.0010

theory. Nyblom and Mäkeläinen (1983) derive the limiting sampling distribution of L when $\hat{e}_i = y_i - \bar{y}$ and $\beta = 0$. It is quite different from that given by Nyblom (1986), where $\hat{e}_i = y_i - \bar{y} - \hat{\beta}(i - (n+1)/2)$. It is likely, although not proved here, that Schmoyer's test has the same limiting distribution as the sampling distribution in the former case, whereas the other two tests have the limiting distribution as in the latter case.

5.4.6 *Conclusions*

The simulation experiments show that the corrected permutation test behaves quite well already when the sample size is in the range 20–36, and when there are few predictor variables. In order to get closer to nominal levels we need larger sample sizes. Schmoyer's alternative test is in most cases conservative, the exceptions being the heteroscedasticity test under heavy-tailed and skew distributions. If the test statistic is such that its distribution depends on the actual predictor values even in large samples, Schmoyer's test may become powerless. Section 5.4.5 gives an example of this possibility. Therefore, I believe the safest approach is to use the corrected permutation test derived here.

5.5 Proofs and Auxiliary Results

For simplicity the dependence on the sample size n is not always shown in the notation, though it is understood that all limits are taken as $n \to \infty$.

Lemma 5.1 *Let λ_{in}, $i = 1, \ldots, n$, $n = 2, 3, \ldots$, be a triangular sequence of real numbers such that (a) $\sum_i \lambda_{in} = 0$ and (b) $\sum_i \lambda_{in}^2 = 1$. If $e_1, e_2, \ldots$ are exchangeable with $Ee_1^6 < \infty$, then*

$$\lim_{n\to\infty} n^{-1/4} \sum_{i=1}^{n} \lambda_{in} e_i = 0, \quad a.s. \tag{5.10}$$

Proof Denote $S_n = \sum_i \lambda_{in} e_i$. Using the assumptions (a) and (b) we get by a direct calculation that

$$ES_n^6 = c_0 + c_1 \sum_i \lambda_{in}^6 + c_2 \sum_i \lambda_{in}^4 + c_3 \left(\sum_i \lambda_{in}^3 \right)^2,$$

where the coefficients c_1, c_2, and c_3 are independent of n. Thus using (b) we get $ES_n^6 \le |c_0| + |c_1| + |c_2| + |c_3| = C$, and consequently

$$E\left(n^{-1/4} S_n\right)^6 \le \frac{C}{n^{3/2}}, \quad \text{for all } n.$$

This implies the convergence of the series $\sum_{n=2}^{\infty} E(S_n/n^{1/4})^6$ which is sufficient for (5.10) by Theorem 1.3.5 of Serfling (1980). □

Lemma 5.2 *Under the assumptions of Lemma 5.1 on* $e_1, e_2, \ldots$

$$\lim_{n\to\infty} \frac{\boldsymbol{e}'\boldsymbol{H}\boldsymbol{e}}{\sqrt{n}} = 0 \quad a.s. \tag{5.11}$$

Proof Let $\boldsymbol{A}$ be a nonsingular square matrix such that $(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{A}\boldsymbol{A}'$. Then Lemma 5.1 is applicable separately to each coordinate of $n^{-1/4}\sum \boldsymbol{A}'\boldsymbol{x}_i e_i = n^{-1/4}\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{e}$. Take its squared norm $n^{-1/2}\boldsymbol{e}'\boldsymbol{X}\boldsymbol{A}\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{e} = n^{-1/2}\boldsymbol{e}'\boldsymbol{H}\boldsymbol{e}$ to prove (5.11). □

By a direct calculation we can prove the following lemma.

Lemma 5.3 *Let* $u_1, \ldots, u_n$ *and* $v_1, \ldots, v_n$ *be two sequences of numbers such that* $\sum u_i = \sum v_i = 0$. *If* $(\pi 1, \ldots, \pi n)$ *is a random permutation of the numbers* $(1, \ldots, n)$ *uniformly distributed over all permutations* Π, *then*

$$E_\pi(u_{\pi 1}^2 v_{\pi 2}^2) = \frac{\left(\sum u_i^2\right)\left(\sum v_i^2\right) - \sum u_i^2 v_i^2}{n(n-1)}, \tag{5.12}$$

$$E_\pi(u_{\pi 1} v_{\pi 1} u_{\pi 2} v_{\pi 2}) = \frac{\left(\sum u_i v_i\right)^2 - \sum u_i^2 v_i^2}{n(n-1)}, \tag{5.13}$$

$$E_\pi(u_{\pi 1}^2 v_{\pi 2} v_{\pi 3}) = E(u_{\pi 1} u_{\pi 2} v_{\pi 3}^2) = \frac{2\sum u_i^2 v_i^2 - \left(\sum u_i^2\right)\left(\sum v_i^2\right)}{n(n-1)(n-2)}, \tag{5.14}$$

$$E_\pi(u_{\pi 1} u_{\pi 2} v_{\pi 2} v_{\pi 3}) = \frac{2\sum u_i^2 v_i^2 - \left(\sum u_i v_i\right)^2}{n(n-1)(n-2)}, \tag{5.15}$$

$$E_\pi(u_{\pi 1} u_{\pi 2} v_{\pi 3} v_{\pi 4}) = \frac{\left(\sum u_i^2\right)\left(\sum v_i^2\right) + 2\left(\sum u_i v_i\right)^2 - 6\left(\sum u_i^2 v_i^2\right)}{n(n-1)(n-2)(n-3)}. \tag{5.16}$$

The formulas (5.12), (5.15), and (5.16) are found also in Schmoyer (1994, p. 1515).

Proof (of Theorem 5.2) Without loss of generality we can assume that the upper bound C for the sum of the squared elements of $\boldsymbol{B}_n = \boldsymbol{B}$ equals to 1. Write $\hat{\boldsymbol{e}} = \boldsymbol{e} - \bar{e}\boldsymbol{1} - \boldsymbol{H}\boldsymbol{e}$ and denote $\boldsymbol{u} = \boldsymbol{e} - \bar{e}\boldsymbol{1}$ and $\boldsymbol{v} = \boldsymbol{H}\boldsymbol{e}$. Then

$$\hat{\boldsymbol{e}}_\pi'\boldsymbol{B}\hat{\boldsymbol{e}}_\pi = \boldsymbol{u}_\pi'\boldsymbol{B}\boldsymbol{u}_\pi - 2\boldsymbol{u}_\pi'\boldsymbol{B}\boldsymbol{v}_\pi + \boldsymbol{v}_\pi'\boldsymbol{B}\boldsymbol{v}_\pi.$$

Clearly, it is sufficient to prove that conditional expectations $E_\pi(\boldsymbol{u}_\pi'\boldsymbol{B}\boldsymbol{v}_\pi)^2$ and $E_\pi|\boldsymbol{v}_\pi'\boldsymbol{B}\boldsymbol{v}_\pi|$ tend to zero a.s.

Let us start with $E_\pi|\boldsymbol{v}_\pi'\boldsymbol{B}\boldsymbol{v}_\pi|$. Denote the eigenvalues and the corresponding eigenvectors of $\boldsymbol{B}$ by λ_k and $\boldsymbol{z}_k$ with $\boldsymbol{z}_k'\boldsymbol{z}_k = 1$. Thus $\boldsymbol{v}_\pi'\boldsymbol{B}\boldsymbol{v}_\pi = \sum \lambda_k(\boldsymbol{z}_k'\boldsymbol{v}_\pi)^2$. Since

$$E_\pi(\boldsymbol{z}_k'\boldsymbol{v}_\pi)^2 \le n^{-1}\sum v_i^2,$$

we find

$$\begin{aligned} E_\pi|\boldsymbol{v}_\pi'\boldsymbol{B}\boldsymbol{v}_\pi| &\le n^{-1}\sum v_i^2 \sum|\lambda_k| \\ &\le n^{-1/2}\boldsymbol{v}'\boldsymbol{v}\left(\sum\lambda_k^2\right)^{1/2} \\ &= n^{-1/2}\boldsymbol{e}'\boldsymbol{H}\boldsymbol{e} = o(1), \text{ a.s.}, \end{aligned} \tag{5.17}$$

by Lemma 5.2 and $\sum\lambda_k^2 = \mathrm{tr}(\boldsymbol{B}^2) = 1$.

Next, let us consider

$$\begin{aligned} E_\pi\left(\sum b_{ii}u_{\pi i}v_{\pi i}\right)^2 &= \left[\frac{\sum u_i^2v_i^2}{n-1} - \frac{(\sum u_iv_i)^2}{n(n-1)}\right]\sum b_{ii}^2 \\ &\quad + \left[\frac{(\sum u_iv_i)^2}{n(n-1)} - \frac{\sum u_i^2v_i^2}{n(n-1)}\right]\left[\sum b_{ii}\right]^2, \end{aligned} \tag{5.18}$$

where we have used Lemma 5.3. By Lemma 5.2

$$n^{-1}\left(\sum u_iv_i\right)^2 = o(1), \quad \text{a.s.}, \tag{5.19}$$

because $\sum u_iv_i = \boldsymbol{e}'\boldsymbol{H}\boldsymbol{e}$. Using the Cauchy-Schwartz inequality we find

$$\begin{aligned} \sum u_i^2v_i^2 &\le \sqrt{\sum u_i^4}\sqrt{\sum v_i^4} \\ &\le \sqrt{\sum u_i^4}\sum v_i^2. \end{aligned}$$

The law of large numbers for exchangeable random variables (Loève 1963, p. 400), gives

$$n^{-1}\sum u_i^4 = n^{-1}\sum(e_i - \bar{e})^4 = O(1), \quad \text{a.s.}$$

Combining this with $\sum v_i^2 = \boldsymbol{e}'\boldsymbol{H}\boldsymbol{e} = o(n^{1/2})$ a.s. (by Lemma 5.2) yields

$$n^{-1}\sum u_i^2v_i^2 = o(1), \quad \text{a.s.} \tag{5.20}$$

By assumption $\sum b_{ii}^2 \le 1$ and hence $(\sum b_{ii})^2 \le n$. Thus (5.18)–(5.20) imply

$$E_\pi \left(\sum b_{ii} u_{\pi i} v_{\pi i} \right)^2 = o(1) \quad \text{a.s.} \tag{5.21}$$

As a second step consider

$$E_\pi \left(\sum b_{ij} u_{\pi i} v_{\pi j} \right)^2 = \sum b_{ij}^2 E_\pi (u_{\pi i}^2 v_{\pi j}^2 + u_{\pi i} v_{\pi i} u_{\pi j} v_{\pi j}) \tag{5.22}$$

$$+ \sum b_{ij} b_{ik} E_\pi (u_{\pi i}^2 v_{\pi j} v_{\pi k} + v_{\pi i}^2 u_{\pi j} u_{\pi k} + 2 u_{\pi i} v_{\pi i} u_{\pi j} v_{\pi k}) \tag{5.23}$$

$$+ \sum b_{ik} b_{jl} E_\pi (u_{\pi i} u_{\pi j} v_{\pi k} v_{\pi l}), \tag{5.24}$$

where the summations are over different subscripts only. Apply Lemma 5.3 to (5.22)–(5.24) and note that the dominant term

$$\left(\sum u_i^2 \right) \left(\sum v_i^2 \right) = \left(\sum (e_i - \bar{e})^2 \right) \boldsymbol{e}' \boldsymbol{H} \boldsymbol{e}$$

is of order $o(n^{3/2})$. Thus, the expectation on the right side of (5.22) is of order $o(n^{-1/2})$, on the line (5.23) of order $o(n^{-3/2})$ and on the line (5.24) of order $o(n^{-5/2})$. Furthermore in (5.22), we have $\sum b_{ij}^2 \le 1$, by assumption. Applications of the Cauchy-Schwartz inequality yield $|\sum b_{ij} b_{ik}| \le n$ in (5.23) and $|\sum b_{ik} b_{jl}| \le n^2$ in (5.24). Then summing up (5.22)–(5.24) we find that

$$E_\pi \left(\sum b_{ij} u_{\pi i} v_{\pi j} \right)^2 = o(n^{-1/2}) \quad \text{a.s.}$$

Combining this with (5.21) we get that $E_\pi (\boldsymbol{u}_\pi' \boldsymbol{B} \boldsymbol{v}_\pi)^2 = o(1)$ a.s. Together with (5.17) this completes the proof. □

Lemma 5.4 *Let $w_1, \ldots, w_n$ and $z_1, \ldots, z_n$ ($n \ge 2$) be two sequences of numbers such that $\sum w_i = \sum z_i = 0$ and $\sum w_i^2 = \sum z_i^2 = 1$. Further, let $(\pi 1, \ldots, \pi n)$ be a random permutation of the numbers $(1, \ldots, n)$ uniformly distributed over all permutations Π. Then*

$$P_\pi \left[\max_{1 \le k \le n} \left| \sum_{j=1}^{k} z_j w_{\pi j} \right| > \varepsilon \right] \le \frac{8}{(n-1)\varepsilon^2}.$$

Proof Denote

$$T_k = \sum_{j=1}^{k} z_j w_{\pi j} \quad \text{and} \quad S_k = \sum_{j=1}^{k} w_{\pi j}.$$

The conditional expectation given $w_{\pi 1}, \ldots, w_{\pi j}$ is denoted by E_j. Using $\sum w_j = 0$ we find for $k > 1$ that

$$E_{k-1}(w_{\pi k}) = -\frac{\sum_{j=1}^{k-1} w_{\pi j}}{n-k+1} = -\frac{S_{k-1}}{n-k+1}.$$

Thus, we can define a martingale sequence as $X_1 = T_1$ and $X_k = T_k + z_k S_{k-1}/(n-k+1)$ for $k = 2, \ldots, n$. Using again the assumption $\sum w_i = 0$ we find that

$$\begin{aligned} X_n = T_n + z_n S_{n-1} &= \sum_{j=1}^{n} z_j w_{\pi j} + z_n \sum_{j=1}^{n-1} w_{\pi j} \\ &= \sum_{j=1}^{n-1} z_j w_{\pi j}. \end{aligned}$$

This implies that $E(X_n^2) = \operatorname{var}(X_n) \leq 1/(n-1)$. Kolmogorov's inequality for submartingales (Billingsley 1979, , p. 414) yields

$$P_\pi(\max_k X_k^2 \geq \eta) \leq \frac{1}{\eta} E(X_n^2) \leq \frac{1}{(n-1)\eta}, \quad \text{for all } \eta > 0. \tag{5.25}$$

An easy calculation shows that

$$E_\pi(S_{k-1}^2) = \frac{(k-1)(n-k)}{n(n-1)}.$$

By this and the inequalities of Bonferroni and Markov we get

$$\begin{aligned} P_\pi\left(\max_{2\leq k\leq n} \frac{z_k^2 S_{k-1}^2}{(n-k+1)^2} \geq \eta\right) &\leq \sum_{k=2}^{n} P_\pi\left(\frac{z_k^2 S_{k-1}^2}{(n-k+1)^2} \geq \eta\right) \\ &\leq \frac{1}{\eta} \sum_{k=2}^{n} E_\pi\left(\frac{z_k^2 S_{k-1}^2}{(n-k+1)^2}\right) \\ &= \frac{1}{\eta} \sum_{k=2}^{n} z_k^2 \frac{(k-1)(n-k)}{(n-k+1)^2 n(n-1)} \\ &\leq \frac{1}{(n-1)\eta} \sum_{k=2}^{n} z_k^2 \\ &\leq \frac{1}{(n-1)\eta}. \end{aligned} \tag{5.26}$$

Since

$$|T_k| \le |X_k| + \frac{|z_k S_{k-1}|}{n-k+1},$$

we finally get by (5.25) and (5.26) that

$$P_\pi(\max|T_k| \ge \varepsilon) \le P_\pi\left(\max|X_k| \ge \frac{\varepsilon}{2}\right) + P_\pi\left(\max\frac{|z_k S_{k-1}|}{n-k+1} \ge \frac{\varepsilon}{2}\right) \le \frac{8}{(n-1)\varepsilon^2}.$$

□

Proof (of Theorem 5.3) An application of Lemma 5.4 with $w_i = v_i/\sqrt{\boldsymbol{v}'\boldsymbol{v}}$ gives

$$\begin{aligned} P_\pi(\max|z_k'\boldsymbol{v}_\pi| \ge \varepsilon) &= P_\pi\left(\max\frac{|z_k'\boldsymbol{v}_\pi|}{\sqrt{\boldsymbol{v}'\boldsymbol{v}}} \ge \frac{\varepsilon}{\sqrt{\boldsymbol{v}'\boldsymbol{v}}}\right) \\ &\le \frac{8\boldsymbol{v}'\boldsymbol{v}}{(n-1)\varepsilon^2} = o(n^{-1/2}),\ \text{a.s.}, \end{aligned} \tag{5.27}$$

for $\boldsymbol{v}'\boldsymbol{v} = \boldsymbol{e}'\boldsymbol{H}\boldsymbol{e} = o(n^{1/2})$ a.s. by Lemma 5.2. Since $z_k'\boldsymbol{Q}(\hat{\boldsymbol{e}}_\pi - \boldsymbol{e}_\pi) = z_k'\boldsymbol{Q}\boldsymbol{v}_\pi = z_k'\boldsymbol{v}_\pi - z_k'\boldsymbol{H}\boldsymbol{v}_\pi$, we get

$$\begin{aligned} P_\pi(\max|z_k'\boldsymbol{Q}(\hat{\boldsymbol{e}}_\pi - \boldsymbol{e}_\pi)| \ge \varepsilon) &\le P_\pi(\max|z_k'\boldsymbol{v}_\pi| \ge \varepsilon/2) \\ &\quad + P_\pi(\max|z_k\boldsymbol{H}\boldsymbol{v}_\pi| \ge \varepsilon/2). \end{aligned} \tag{5.28}$$

Since $\boldsymbol{H}$ is idempotent, we get by the Cauchy-Schwartz inequality that

$$(z_k'\boldsymbol{H}\boldsymbol{v}_\pi)^2 \le (z_k'\boldsymbol{H}z_k)(\boldsymbol{v}_\pi'\boldsymbol{H}\boldsymbol{v}_\pi).$$

The idempotency also implies that $z_k'\boldsymbol{H}z_k \le z_k'z_k \le 1$ for each k. Further,

$$E_\pi(\boldsymbol{v}_\pi'\boldsymbol{H}\boldsymbol{v}_\pi) = (\mathrm{tr}\boldsymbol{H})\boldsymbol{v}'\boldsymbol{v}/(n-1) = q\boldsymbol{e}'\boldsymbol{H}\boldsymbol{e}/(n-1) = o(n^{1/2}) \quad \text{a.s.}$$

by (5.11). Together with (5.27) and (5.28) this yields

$$P_\pi(\max|z_k'\boldsymbol{Q}(\hat{\boldsymbol{e}}_\pi - e_\pi)| > \varepsilon) \longrightarrow 0, \quad \text{a.s.}$$

for each $\varepsilon > 0$. The proof is completed by noting that if $n > M$

$$\begin{aligned} \max_{m_0 \le k \le m_1} \frac{|z_k'\boldsymbol{Q}(\hat{\boldsymbol{e}}_\pi - \boldsymbol{e}_\pi)|}{\sqrt{z_k'\boldsymbol{Q}z_k}} &\le \frac{\max_{m_0 \le k \le m_1}|z_k'\boldsymbol{Q}(\hat{\boldsymbol{e}}_\pi - \boldsymbol{e}_\pi)|}{\sqrt{\min_{m_0 \le k \le m_1} z_k'\boldsymbol{Q}z_k}} \\ &\le \frac{\max_{m_0 \le k \le m_1}|z_k'\boldsymbol{Q}(\hat{\boldsymbol{e}}_\pi - \boldsymbol{e}_\pi)|}{\sqrt{b}}. \end{aligned}$$

□

Acknowledgements The author thanks two anonymous referees for their useful comments.

References

Anderson, M.J., Robinson, J.: Permutation tests for linear models. Aust. N. Z. J. Stat. **43**, 75–88 (2001)
Anderson, M.J., ter Braak, C.J.F.: Permutation tests for multi-factorial analysis of variance. J. Stat. Comput. Simul. **73**, 85–113 (2003)
Billingsley, P.: Probability and Measure. Wiley, New York (1979)
Chen, J.: Testing for a change point in linear regression models. Commun. Stat. Theory Methods **27**, 2481–2493 (1998)
Cook, R.D., Weisberg, S.: Residuals and Influence in Regression. Chapman and Hall, New York (1982)
Cook, R.D., Weisberg, S.: Diagnostics for heteroscedasticity in regression. Biometrika **70**, 1–10 (1983)
Edgington, E.S.: Randomization Tests, 3rd edn. Marcel Dekker, New York (1995)
Efron, B.: Bootstrap methods: another look at the jacknife. Ann. Stat. **7**, 1–26 (1979)
Fisher, R.A.: Design of Experiments. Oliver and Boyd, Edinburgh (1935)
Freedman, D., Lane, D.: A nonstochastic interpretation of reported significance levels. J. Bus. Econ. Stat. **1**, 292–298 (1983)
Good, P.: Permutation Tests, 3rd edn. Springer, New York (2005)
Huh, M.-H., Jhun, M.: Random permutation testing in multiple linear regression. Commun. Stat. Theory Methods **30**, 2023–2032 (2001)
Holbert, D.: A Bayesian analysis of a switching linear model. J. Econ. **19**, 77–87 (1982)
Imhof, J.P.: Computing the distribution of quadratic forms in normal variables. Biometrika **48**, 419–426 (1961)
Kempthorne, O.: Randomization–II. In: Kotz, S., Johnson, N.L. (eds.) Encyclopedia of Statistical Sciences, vol. 7, pp. 519–524. Wiley, New York (1986)
Kennedy, P.E.: Randomization tests in econometrics. J. Bus. Econ. Stat. **13** 85–94 (1995)
Kim, H.-J., Siegmund, D.: The likelihood ratio test for a change-point in simple linear regression. Biometrika **76**, 409–423 (1989)
Kim, H.-J., Fay, M.P., Feuer, E.J., Midthune, D.: Permutation tests for jointpoint regression with applications to cancer rates. Stat. Med. **19**, 335–351 (2000)
Koutsoyiannis, A.: Theory of Econometrics. Macmillan, London (1973)
Lehmann, E.L., Romano, J.P.: Testing Statistical Hypotheses, 3rd edn. Chapman and Hall, New York (2005)
LePage, R., Podgórski, K.: Resampling permutations in regression without second moments. J. Multivar. Anal. **57**, 119–141 (1996)
Loève, M.: Probability Theory. Van Nostrand, New York (1963)
Manly, B.F.J.: Randomization, Bootstrap, and Monte Carlo Methods in Biology, 2nd edn. Chapman and Hall, London (1997)
Nyblom, J.: Testing for deterministic linear trend in time series. J. Am. Stat. Assoc. **81**, 545–549 (1986)
Nyblom, J., Mäkeläinen, T.: Comparisons of tests for the presence of random walk coefficients in a simple linear model. J. Am. Stat. Assoc. **78**, 856–864 (1983)
Oja, H.: On permutation tests in multiple regression and analysis of covariance problems. Aust. J. Stat. **29**, 91–100 (1987)
Pesarin, F.: Multivariate Permutation Tests. Wiley, New York (2001)
Pitman, E.J.G.: Significance tests which may be applied to samples from any populations. J. R. Stat. Soc. B **4**, 119–130 (1937a)

Pitman, E.J.G.: Significance tests which may be applied to samples from any populations. II. The correlation coefficient. J. R. Stat. Soc. B **4**, 225–232 (1937b)
Pitman, E.J.G.: Significance tests which may be applied to samples from any populations. III. The analysis of variance test. Biometrika **29**, 322–335 (1938)
Ryan, T., Joiner, B., Ryan, B.: Minitab Student Handbook. Duxbury, North Scituate (1976)
Rao, C.R.: Linear Statistical Inference and Its Applications, 2nd edn. Wiley, New York (1973)
Schmoyer, R.L.: Permutation tests for correlation in regression errors. J. Am. Stat. Assoc. **89**, 1507–1516 (1994)
Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
Serfling, R.J.: Approximation Theorems of Mathematical Statistics. Wiley, New York (1980)

Chapter 6
Highly Robust and Highly Finite Sample Efficient Estimators for the Linear Model

Ezequiel Smucler and Víctor J. Yohai

Abstract In this paper, we propose a new family of robust regression estimators, which we call bounded residual scale estimators (BRS-estimators) which are simultaneously highly robust and highly efficient for small samples with normally distributed errors. To define these estimators it is required to have a robust M-scale and a family of robust MM-estimators. We start by choosing in this family a highly robust initial estimator but not necessarily highly efficient. Loosely speaking, the BRS-estimator is defined as the estimator in the MM family which is closest to the LSE among those with a robust M-scale sufficiently close to the one of the initial estimators. The efficiency of the BRS is derived from the fact that when there are not outliers in the sample and the errors are normally distributed, the scale of the LSE is similar to the one of the initial estimator. The robustness of the BRS-estimator comes from the fact that its robust scale is close to the one of the initial highly robust estimator. The results of a Monte Carlo study show that the proposed estimator has a high finite-sample efficiency, and is highly resistant to outlier contamination.

Keywords Brakdown point • Finite sample efficiency • MM-estimators

6.1 Introduction

Robust statistics aims at providing methods for statistical modelling which are reliable even in the presence of atypical data or outliers. The main objective of robust estimation is to obtain estimators which are almost as good as the optimal classical estimator (say the maximum likelihood estimator) when the data

E. Smucler
Instituto de Cálculo, Universidad de Buenos Aires, Intendente Güiraldes 2160, 1428 Buenos Aires, Argentina
e-mail: esmucler@dm.uba.ar

V.J. Yohai (✉)
Departamento de Matemáticas and Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Intendente Güiraldes 2160, 1428 Buenos Aires, Argentina
e-mail: vyohai@dm.uba.ar

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_6

distribution coincides with the nominal one and that are not much affected by outlier contamination.

In this paper, we are concerned with robust estimation for the linear model. We assume that we observe $(\mathbf{x}_i^{\mathrm{T}}, y_i)$ $i = 1, \ldots, n$, i.i.d. $(p+1)$-dimensional vectors, where y_i is the response variable and $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of random carriers, satisfying

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + u_i \text{ for } i = 1, \ldots, n, \tag{6.1}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the vector of regression coefficients to be estimated and u_i is independent of $\mathbf{x}_i$.

Let F_0 be the distribution of the errors u_i, G_0 the distribution of the carriers $\mathbf{x}_i$ and H_0 the distribution of $(\mathbf{x}_i^{\mathrm{T}}, y_i)$. Then H_0 satisfies

$$H_0(\mathbf{x}, y) = G_0(\mathbf{x}) F_0(y - \mathbf{x}^{\mathrm{T}} \boldsymbol{\beta}_0). \tag{6.2}$$

The least-squares estimator (LSE) of $\boldsymbol{\beta}_0$ is defined by

$$\hat{\boldsymbol{\beta}}_{LSE} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}),$$

where $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}$.

When the errors are normally distributed, the LSE of $\boldsymbol{\beta}_0$ coincides with the maximum likelihood estimator (MLE), and has minimum variance among unbiased estimators. However it is a well-known fact that the LSE is not robust, that is, it is very sensitive to small deviations from the model assumptions. Moreover, it has zero breakdown point, that is, a single sufficiently large outlying observation can completely spoil the LSE.

The asymptotic efficiency of a regression estimator is defined as the ratio between the MLE asymptotic variance and the estimator's asymptotic variance. However, when the sample size is not very large, the efficiency for this sample size may turn out to be much lower than the asymptotic efficiency. Given a positive integer n, the efficiency of an estimator when the sample size is equal n is defined as the ratio between the mean squared errors (MSE) of the MLE and that of the estimator in question for that sample size. It is clear that for practical purposes, the efficiency for finite sample size is the relevant measure of efficiency.

The first regression equivariant estimator to achieve the optimal 1/2 asymptotic breakdown point was the least median of the squares estimator (LMSE), first proposed by Hampel (1975) and further developed by Rousseeuw (1984). The LMSE converges at a $n^{-1/3}$ rate and thus its relative efficiency with respect to the LSE is 0 (see Davies 1990).

S-estimators, introduced in Rousseeuw and Yohai (1984), were the first regression estimators to combine both: 1/2 asymptotic breakdown point and $n^{-1/2}$ rate of convergence. However, Hössjer (1992) proved that any regression S-estimator

calculated using a smooth loss function, and tuned to have 1/2 asymptotic breakdown point for regression equivariant estimators, is doomed to have an asymptotic efficiency at the normal distribution of at most 0.329.

MM-estimators, introduced in Yohai (1987), are regression estimators that can be tuned to attain both a high breakdown point and an asymptotic efficiency at the normal distribution as close to one as desired.

In Gervini and Yohai (2002), the authors introduced the robust and efficient weighted least squares estimators (REWLSE), which combine a high breakdown point with full asymptotic efficiency for normal errors.

However, as will be seen in the simulation study we report, when the sample size is not very large, both the REWLSE and the MM-estimator can have a finite-sample efficiency that is much lower than their asymptotic one.

On the other hand, an estimator that has a high breakdown point can still be largely affected by a small fraction of contaminated observations. Bondell and Stefanski (2013) propose a high breakdown point regression estimator with high finite sample normal efficiency. However, we show in Sect. 6.4 that this estimator is highly unstable under outlier contamination.

In Maronna and Yohai (2014), the authors propose a family of estimators, called distance constrained maximum likelihood estimators (DCML-estimators), which are simultaneously robust and highly efficient for small samples.

In this paper, we propose a new family of robust regression estimators, which we call bounded residual scale estimators (BRS-estimators). These estimators are simultaneously highly robust and highly efficient for small samples with normally distributed errors. To define these estimators it is required to have a robust M-scale and a family of robust MM-estimators. We start by choosing in this family a highly robust initial estimator but not necessarily highly efficient. Loosely speaking, the BRS-estimator is defined as the estimator in the MM family which is closest to the LSE among those with a robust M-scale sufficiently close to the one of the initial estimator. The efficiency of the BRS-estimator is derived from the fact that when there are not outliers in the sample and the errors are normally distributed, the scale of the LSE is similar to the one of the initial estimator. The robustness of the BRS-estimator is due to the fact that its robust scale is close to the one of the initial highly robust estimator. The results of a Monte Carlo study show that the proposed estimator has a high finite-sample efficiency and is highly resistant to outlier contamination.

In Sect. 6.2, we review the definition and some of the most important properties of MM- and S-estimators of regression. In Sect. 6.3 we introduce the BRS-estimators and study their robust and asymptotic behaviour. In Sect. 6.4 we report the results of simulation study and apply the proposed estimator to a published data set. Conclusions and possible further extensions of the proposed estimators to other models are provided in Sect. 6.5.

6.2 S-Estimators and MM-Estimators

6.2.1 M-Estimates of Scale and S-Estimators

(A) Let ρ be a real valued function satisfying:

- $\rho(x)$ is a non-decreasing, continuous and even function.
- $\rho(0) = 0$.
- $\sup \rho(x) = 1$, and if $\rho(u) < 1$ and $0 \leq u < v$ then $\rho(u) < \rho(v)$.

Any function that satisfies the assumptions in (A) will be called a ρ-function.

Let $0 < b < 1$, and let F be a distribution function. Let ρ be a ρ-function. Then, following Huber (1981), we define the M-scale functional as

$$s(F) = \inf \left\{ s \geq 0 : E_F \rho \left(\frac{u}{s} \right) \leq b \right\}.$$

It is easy to show that $s(F) > 0$ if and only if $P_F(u = 0) < 1 - b$, and in this case

$$E_F \rho \left(\frac{u}{s(F_0)} \right) = b.$$

Given a sample $\mathbf{u} = (u_1, \ldots, u_n)$ from F, the corresponding M-estimate of scale $s_n(\mathbf{u})$ is defined by

$$s_n(\mathbf{u}) = \inf \left\{ s \geq 0 : \frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{u}{s_n(\mathbf{u})} \right) \leq b \right\}.$$

It is easy to prove that $s_n(\mathbf{u}) > 0$ if and only if $\#\{i : u_i = 0\} < (1 - b)n$, and in this case

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{u}{s_n(\mathbf{u})} \right) = b. \tag{6.3}$$

It can be readily verified that scale M-estimates are scale equivariant. Huber (1981) proves that the asymptotic breakdown point of an M-estimate of scale is given by $\min\{b, 1 - b\}$. Taking $b = 1/2$ gives an asymptotic breakdown point of 1/2, the maximum possible for scale equivariant estimates of scale.

Given a sample $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, \ldots, n$ from the model given in (6.2), Rousseeuw and Yohai (1984) define the S-estimator of regression as

$$\hat{\boldsymbol{\beta}}_{S,n} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} s_n(\mathbf{r}(\boldsymbol{\beta})).$$

S-estimators of regression have been shown to be strongly consistent for $\boldsymbol{\beta}_0$ under very general conditions. See Fasano et al. (2012) for details. One measure of the robustness of an estimator introduced by Donoho and Huber (1983) is the finite-sample replacement breakdown point. In the case of regression this measure is defined as follows. Given a sample $\mathbf{z}_i = (\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, \ldots, n$, let $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ and let $\hat{\boldsymbol{\beta}}(\mathbf{Z})$ be a regression estimator. The finite-sample replacement breakdown point of $\hat{\boldsymbol{\beta}}$ is then defined as

$$FBP(\hat{\boldsymbol{\beta}}) = \frac{m^*}{n},$$

where

$$m^* = \max\{m \geq 0 : \hat{\boldsymbol{\beta}}(\mathbf{Z}_m) \text{ is bounded for all } \mathbf{Z}_m \in \mathscr{Z}_m\},$$

and $\mathscr{Z}_m$ is the set of all datasets with at least $n - m$ elements in common with $\mathbf{Z}$.

S-estimators can always be tuned so as to attain the maximum possible finite-sample replacement breakdown point for regression equivariant estimators, which is given by

$$\frac{1}{n}\left[\frac{n - k^*(\mathbf{X}) - 1}{2}\right], \tag{6.4}$$

where $k^*(\mathbf{X})$ is the maximum number of observations $\mathbf{x}_i, \leq 1 \leq i \leq n$ lying in a subspace and [ ] denotes integer part. If the sample is in general position, that is, if $k^*(\mathbf{X}) = p - 1$, (6.4) becomes

$$\frac{1}{n}\left[\frac{n - p}{2}\right],$$

which is approximately 1/2 for large n. See Maronna et al. (2006) for details. Recall, however, that S-estimators cannot combine high breakdown point with high efficiency at the normal distribution.

6.2.2 MM-Estimators

Let $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, \ldots, n$, be a sample satisfying (6.2), then Yohai (1987) defines the MM-estimator of regression in three stages as follows:

1. Compute $\hat{\boldsymbol{\beta}}_{0,n}$, a high breakdown point estimator of $\boldsymbol{\beta}_0$.
2. Let ρ_0 be a ρ-function, and let $0 < b < 1$. Compute the corresponding M-estimate of scale of the residuals of $\hat{\boldsymbol{\beta}}_{0,n}$.

3. Let ρ_1 be another ρ-function, such that $\rho_1 \leq \rho_0$. Let $\psi_1 = \rho_1'$. The MM-estimator of regression $\hat{\boldsymbol{\beta}}_{MM,n}$ is defined as any solution of

$$\sum_{i=1}^{n} \mathbf{x}_i \psi_1 \left(\frac{r_i(\boldsymbol{\beta})}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{0,n}))} \right) = 0$$

satisfying

$$L(\hat{\boldsymbol{\beta}}_{MM,n}) \leq L(\hat{\boldsymbol{\beta}}_{0,n}),$$

where

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_1 \left(\frac{r_i(\boldsymbol{\beta})}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{0,n}))} \right),$$

and putting $\rho_1\left(\frac{0}{0}\right) = 0$.

Yohai (1987) proves that under very general conditions, MM-estimators are strongly consistent for $\boldsymbol{\beta}_0$, and furthermore

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{MM,n} - \boldsymbol{\beta}_0) \rightarrow^d N_p(\mathbf{0}, \sigma_0^2 v(\psi_1, F_0) \mathbf{V}_{\mathbf{x}}^{-1}), \tag{6.5}$$

where $\mathbf{V}_{\mathbf{x}} = E_{G_0}(\mathbf{x}\mathbf{x}^{\mathrm{T}})$, $\sigma_0 = s(F_0)$, where the M-scale s is defined using ρ_0 and

$$v(\psi, F_0) = \frac{E_{F_0} \psi\left(u/\sigma_0\right)^2}{\left(E_{F_0} \psi'\left(u/\sigma_0\right)\right)^2}.$$

Besides, he shows that ρ_1 can be chosen so that the resulting MM-estimator has simultaneously the two following properties:

- Normal asymptotic efficiency as close to one as desired.
- Breakdown point equal or larger than that of the initial estimator.

Maronna et al. (2006) recommend the use of MM-estimators using an S-estimator with high breakdown point as an initial estimator and $\rho_0 = \rho^B(u/c_0)$, where ρ^B is Tukey's bisquare loss function, given by

$$\rho^B(u) = 1 - I(|u| \leq 1)(1 - (u)^2)^3$$

and c_0 is a tuning parameter. This parameter should be chosen so that the resulting M-estimate of scale be consistent in the case of normal errors. The function ρ_1 can be taken equal to $\rho^B(u/c_1)$ where $c_1 \geq c_0$. Maronna et al. (2006) recommend to choose c_1 so that the MM-estimator has an asymptotic efficiency of 85 % at the normal distribution. The reason for choosing an 85 % asymptotic efficiency at the normal distribution is that at this level of the efficiency the MM-estimator has the

same maximum asymptotic bias as the initial S-estimator of regression for the case of normal errors and normal carriers.

6.3 BRS-Estimators

In order to define the proposed estimators, we first introduce some notation. Let ρ be a ρ-function, and for each $c > 0$, let $\rho_c(u) = \rho(u/c)$. Fix c_0 and $0 < b < 1$, and let s_n be the M-estimate of scale defined by (6.3) using ρ_{c_0} as ρ and b. Let $\hat{\boldsymbol{\beta}}_{S,n}$ be the S-estimator which minimizes $s_n(\mathbf{r}(\boldsymbol{\beta}))$. Consider a sequence $(\delta_n)_n$, and constants c_1, c_2 satisfying $\delta_n \geq 0$ and $c_0 \leq c_1 \leq c_2$. Let $\hat{\boldsymbol{\beta}}_{c_1,n}$ be the MM-estimator calculated using $\hat{\boldsymbol{\beta}}_{S,n}$ as initial estimator, the scale $s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{S,n}))$ and ρ_{c_1} as ρ_1. For each $c \in (c_1, c_2]$, let $\hat{\boldsymbol{\beta}}_{c,n}$ be the MM-estimator calculated using $\hat{\boldsymbol{\beta}}_{c_1,n}$ as initial estimator, the scale $s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{c_1,n}))$ and ρ_c as ρ_1. Then given a sample $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, \ldots, n$, from (6.2) define

$$c^* = \sup\{c \in [c_1, c_2] : s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{c,n})) \leq (1 + \delta_n) s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{c_1,n}))\}. \tag{6.6}$$

We define the BRS-estimator, $\hat{\boldsymbol{\beta}}_{BRS,n}$ as

$$\hat{\boldsymbol{\beta}}_{BRS,n} = \hat{\boldsymbol{\beta}}_{c^*,n}.$$

Note that $s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{BRS,n})) \leq (1 + \delta_n) s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{c_1,n}))$. It is easy to see that BRS-estimators are regression, affine and scale equivariant.

We recommend the BRS-estimator with ρ in Tukey's bisquare family, $b = 0.5(1 - p/n)$, $c_0 = 1.54$, $c_1 = 3.44$, $c_2 = 7.14$, and $\delta_n = 0.8p/n$. The value of c_0 makes the scale estimate consistent in the case of normal errors and the value of b makes the S- and MM-estimators have maximum FBP when the sample is in general position. The asymptotic efficiency of the MM-estimators with tuning parameters c_1 and c_2 is equal to 85 % and 99 % respectively. The values for δ_n were obtained by trial and error after several Monte Carlo experiments with different values of p and n. As we will see in Sect. 6.4, this estimator is simultaneously highly efficient and highly robust.

The following theorem implies that, under quite general conditions, a BRS-estimator has a finite-sample replacement breakdown point that is at least as high as that of the initial MM-estimator $\hat{\boldsymbol{\beta}}_{c_1,n}$.

Theorem 6.1 *Let s_n be an M-estimate of scale. Let $\hat{\boldsymbol{\beta}}_{0,n}$ and $\hat{\boldsymbol{\beta}}_{1,n}$ be regression estimators. Suppose that*

(i) For some $K > 0$

$$s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{1,n})) \leq K s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{0,n}))$$

for all $n \in \mathbb{N}$.

(ii)

$$FBP(\hat{\boldsymbol{\beta}}_{0,n}) < b.$$

(iii)

$$[nb] \leq \left[\frac{n - k^*(\mathbf{X}) - 1}{2}\right],$$

where b is the right-hand side of (6.3).

Then

$$FBP(\hat{\boldsymbol{\beta}}_{1,n}) \geq FBP(\hat{\boldsymbol{\beta}}_{0,n}).$$

Proof Consider observations $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, \ldots, n$ and let

$$m < nFBP(\hat{\boldsymbol{\beta}}_{0,n}).$$

Take $C \subset \{1, 2, \ldots, n\}$ such that $\#C = m$ and a sequence $(\mathbf{x}_{N,i}^{\mathrm{T}}, y_{N,i})_N$ such that $(\mathbf{x}_{N,i}^{\mathrm{T}}, y_{N,i}) = (\mathbf{x}_i^T, y_i)$ for $i \notin C$ and all $N \in \mathbb{N}$. Let $\hat{\boldsymbol{\beta}}_{0,n}^N$ and $\hat{\boldsymbol{\beta}}_{1,n}^N$ denote the estimators $\hat{\boldsymbol{\beta}}_{0,n}$ and $\hat{\boldsymbol{\beta}}_{1,n}$ computed in $(\mathbf{x}_{N,i}^{\mathrm{T}}, y_{N,i})_N$. Since there are a finite number of sets included in $\{1, \ldots, n\}$, to prove the theorem it will be enough to show that $(\hat{\boldsymbol{\beta}}_{1,n}^N)_N$ is bounded. Suppose that this is not true, then passing eventually to a subsequence we can assume that $\|\hat{\boldsymbol{\beta}}_{1,n}^N\| \to \infty$ when $N \to \infty$ and that there exists $\boldsymbol{\beta}^*$ such that

$$\lim_{N \to \infty} \hat{\boldsymbol{\beta}}_{1,n}^N / \|\hat{\boldsymbol{\beta}}_{1,n}^N\| = \boldsymbol{\beta}^*.$$

Let $D = \{1, 2, \ldots n\} - C$ and let

$$D^* = \{i : i \in D \text{ and } \boldsymbol{\beta}^{*'}\mathbf{x}_i \neq 0\}.$$

Then

$$\#D^* \geq n - m - k^*(\mathbf{X}).$$

Since

$$n - k^*(\mathbf{X}) \geq 2\left[\frac{n - k^*(\mathbf{X}) - 1}{2}\right] + 1$$

and

$$m \leq [nb] \leq \left[\frac{n - k^*(\mathbf{X}) - 1}{2}\right],$$

we have that

$$\#D^* \geq n - m - k^*(\mathbf{X}) \geq 1 + \left[\frac{n - k^*(\mathbf{X}) - 1}{2}\right] \geq 1 + [nb] > nb.$$

It is immediate that for $i \in D^*$ we have $r_{N,i}(\hat{\boldsymbol{\beta}}^N_{1,n}) \to \infty$. Therefore by Lemma 5.2 of Maronna et al. (2006), $s_n(\mathbf{r}_N(\hat{\boldsymbol{\beta}}^N_{1,n})) \to \infty$. On the other hand, since $(\hat{\boldsymbol{\beta}}^N_{0,n})_N$ is bounded, we have that there exists K such that $\left|r_{N,i}(\hat{\boldsymbol{\beta}}^N_{0,n})\right| \leq K$ for all $i \in D$. Since $m < nb$, we have $\#D > n - nb$ and then by Lemma 5.2 of Maronna et al. (2006), $s_n(\mathbf{r}_N(\hat{\boldsymbol{\beta}}^N_{0,n}))$ is bounded, contradicting assumption (i) of the theorem. □

Remark 6.1 Suppose that $b = ([(n - k^*(\mathbf{X}) - 1)/2] + \gamma)/n$ for some $\gamma \in (0, 1)$ in (6.3), then the MM-estimator that uses the S-estimator that minimizes $s_n(r(\boldsymbol{\beta}))$ as initial estimator has maximal breakdown point. In this case, Theorem 6.1 implies that $\hat{\boldsymbol{\beta}}_{BRS,n}$ also has maximal breakdown point.

In order to prove the consistency of BRS-estimators, the following additional assumptions are needed:

(B1) The ρ function used to define the BRS-estimator is eventually constant.
(B2) $P_{G_0}\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} = 0\right) < 1 - b$ for all non-zero $\boldsymbol{\beta} \in \mathbb{R}^p$, where b was used to define s_n.
(B3) F_0 has an even continuous density, f_0, that is a monotone decreasing function of $|u|$ and a strictly decreasing function of $|u|$ in a neighbourhood of 0.

Consistency of BRS-estimators is an immediate corollary of the following Theorem.

Theorem 6.2 *Let $(\mathbf{x}_i^T, y_i)$, $i = 1, \ldots, n$, be i.i.d observations with distribution H_0, which satisfies (6.2). Assume (B1), (B2), and (B3) hold. Let $(\hat{\boldsymbol{\beta}}_{1,n})_n$ be a sequence of regression equivariant estimators. Let s_n be an M-estimate of scale. Suppose that*

$$s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{1,n})) \to s(F_0) \; a.s..$$

Then $(\hat{\boldsymbol{\beta}}_{1,n})_n$ is strongly consistent for $\boldsymbol{\beta}_0$.

Proof By the regression equivariance of $\hat{\boldsymbol{\beta}}_{1,n}$, we can assume that $\boldsymbol{\beta}_0 = \mathbf{0}$. We will show first that $(\hat{\boldsymbol{\beta}}_{1,n})_n$ is bounded with probability one. Let

$$A = \left\{s_n(\hat{\boldsymbol{\beta}}_{1,n}) \to s(F_0)\right\}.$$

By the assumptions of the theorem we have

$$P(A) = 1.$$

By Lemma 4.6 of Yohai and Zamar (1986) there exists K such that

$$\lim_{n\to\infty} \inf_{\|\boldsymbol{\beta}\|\geq K} s_n(\mathbf{r}(\boldsymbol{\beta})) \geq s(F_0) + 1 \text{ a.s..}$$

Let

$$B = \left\{ \lim_{n\to\infty} \inf_{\|\boldsymbol{\beta}\|\geq K} s_n(\mathbf{r}(\boldsymbol{\beta})) \geq s(F_0) + 1 \right\}$$

and put $D = A \cap B$, then clearly $P(D) = 1$. We will first show that the samples in D are bounded. Suppose that there exists a sample in D which is not bounded. Then for this sample $\limsup_{n\to\infty} s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{1,n})) \geq s(F_0)+1$, contradicting the fact that $\lim_{n\to\infty} s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{1,n})) = s(F_0)$. For any positive integer j, define E_j as

$$E_j = \left\{ \lim_{n\to\infty} \sup_{||\boldsymbol{\beta}||\leq j} |s_n(\mathbf{r}(\boldsymbol{\beta})) - s(\boldsymbol{\beta}, H_0)| = 0 \right\}.$$

By Lemma 4.5 of Yohai and Zamar (1986), $P(E_j) = 1$ for all j. Put $E = (\cap_{j=1}^{\infty} E_j) \cap D$, then $P(E) = 1$ too. We will show that for any sample in E we have

$$\lim_{n\to\infty} \hat{\boldsymbol{\beta}}_{1,n} = \mathbf{0}. \tag{6.7}$$

Suppose that there exists a sample in D that does not satisfy (6.7). Then since this sample is bounded, there exists a subsequence $(n_i)_i$ and $\boldsymbol{\beta}^* \neq \mathbf{0}$ such that

$$\lim_{i\to\infty} \hat{\boldsymbol{\beta}}_{1,n_i} = \boldsymbol{\beta}^*.$$

Take j_0 such that $||\boldsymbol{\beta}^*|| < j_0$ and let $a = ||\boldsymbol{\beta}^*||/2$. Then there exists i_0 such that for $i \geq i_0$ we have $a \leq ||\widehat{\boldsymbol{\beta}}_{1,n_i}|| \leq j_0$. Then, since the sample belongs to E_{j_0}

$$\limsup_{i\to\infty} \left(\left| s_{n_i}(\mathbf{r}(\hat{\boldsymbol{\beta}}_{1,n_i})) - s(\hat{\boldsymbol{\beta}}_{1,n_i}, H_0) \right| \right) \leq \lim_{i\to\infty} \sup_{|\boldsymbol{\beta}||\leq j_0} (|s_{n_i}(\mathbf{r}(\boldsymbol{\beta})) - s(\boldsymbol{\beta}, H_0)|) = 0.$$

Then, using Lemma 4.1 of Yohai and Zamar (1986) we have

$$\begin{aligned} \liminf_{i\to\infty} s_{n_i}(\mathbf{r}(\hat{\boldsymbol{\beta}}_{1,n_i})) &= \liminf_{i\to\infty} s(\hat{\boldsymbol{\beta}}_{1,n_i}, H_0) \\ &\geq \inf_{a\leq||\boldsymbol{\beta}||\leq j_0} s(\boldsymbol{\beta}, H_0) \\ &> s(F_0). \end{aligned}$$

This contradicts the fact that we are dealing with a sample belonging to A. □

Corollary 6.1 *Under the assumptions of Theorem 6.2, if* $(\hat{\boldsymbol{\beta}}_{S,n})_n$ *and* $(\hat{\boldsymbol{\beta}}_{c_1,n})_n$ *are strongly consistent for* $\boldsymbol{\beta}_0$*, and if* $\delta_n \to 0$*, then* $(\hat{\boldsymbol{\beta}}_{BRS,n})_n$ *is strongly consistent for* $\boldsymbol{\beta}_0$.

Even though we were not able to derive the asymptotic distribution of BRS-estimators, we propose the following ad-hoc procedure for approximate asymptotic inference on a BRS-estimator. Given a sample $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, \ldots, n$, we use the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{c^*,n}$, that is, the asymptotic distribution of the MM-estimator with tuning constant equal to c^*, as an approximation of the distribution of the BRS-estimator.

When the BRS-estimator coincides with $\hat{\boldsymbol{\beta}}_{c^*,n}$, according to (6.5), the approximate covariance matrix of the BRS-estimator is $\sigma_0^2 v(\psi_{c^*}, F_0)\mathbf{V}_{\mathbf{x}}{}^{-1}$. In this case σ_0 can be estimated by $s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{c^*,n}))$, $v(\psi_{c^*}, F_0)$ by $v(\psi_{c^*}, F_n)$ where F_n is the residuals empirical distribution and $\mathbf{V}_{\mathbf{x}}$, as proposed in Stahel et al. (1991), by the robust estimator

$$\hat{\mathbf{V}}_{\mathbf{x}} = \sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}},$$

where

$$w_i = \frac{w(r_i(\hat{\boldsymbol{\beta}}_{c^*,n})/s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{c^*,n}))}{\sum_{j=1}^{n} w(r_j(\hat{\boldsymbol{\beta}}_{c^*,n})/s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{c^*,n}))},$$

and

$$w(u) = \frac{\psi_{c^*}(u)}{u}.$$

Although this procedure is not strictly justified, it seems to work quite well in practice. To evaluate the accuracy of this approximation in the case of the BRS-estimator described above Theorem 6.1, we computed the average length and coverage probabilities for the confidence intervals with nominal 95 % asymptotic confidence level for each of the regression parameters, for a model with $p = 5$ variables and an intercept. We consider two possible error distributions: normal and Student's t-distribution with three degrees of freedom (t_3). The values were averaged over the six parameters, and were computed through a simulation with 3000 replicates. Results are shown in Table 6.1.

It is seen that the approximation works fairly well for $n \geq 50$. These results are similar to those obtained by Bondell and Stefanski (2013). Note that when $n = 20$ we have $n/p = 3.33$ and according to the usual rule of thumb it is recommended to have $n/p \geq 5$.

Table 6.1 Mean interval lengths and coverage probabilities for intervals for the recommended BRS-estimator with nominal 95 % asymptotic confidence level, for a model with $p = 5$ parameters plus intercept, averaged over the six parameters

	Normal		t_3	
n	Length	Coverage	Length	Coverage
20	0.855	86.4 %	1.17	86.3 %
50	0.555	92.7 %	0.730	92.8 %
100	0.393	94.1 %	0.514	94.3 %
200	0.277	94.2 %	0.362	95.2 %
500	0.176	94.3 %	0.228	94.2 %

6.4 Simulations

We generated $(\mathbf{x}_1^{\mathrm{T}}, y_1), \ldots, (\mathbf{x}_n^{\mathrm{T}}, y_n)$ i.i.d observations according to the model (6.1). Since all the estimators considered are regression equivariant, we take without loss of generality $\boldsymbol{\beta}_0 = 0$. We took $p = 5, 10, 20$ and $n = Kp$ with $K = 5, 10, 20$. The total number of Monte Carlo replications was $N_{rep} = 1000$ in all cases.

We compared the robustness and finite-sample efficiency of the following estimators:

- An MM-estimator with 85 % asymptotic efficiency, henceforth MM85, with ρ_1 in Tukey's bisquare family. This is the recommended estimator in Maronna et al. (2006).
- The Bondell and Stefanski (2013) estimator, henceforth B-S.
- The REWLSE of Gervini and Yohai (2002).
- The BRS-estimator described above Theorem 6.1.
- The DCML-estimator proposed in Maronna and Yohai (2014).
- An S-estimator based on a ρ-function in Tukey's bisquare family.

MM- and S-estimators were computed using the functions `lmrob..M..fit` and `lmrob.S` in the *robustbase* R package. The parameter c^* defined in (6.6) was computed via a grid search. The code to calculate B-S, REWLSE and DCML was kindly provided by the authors.

6.4.1 Efficiency

To assess the finite-sample efficiency of the estimators when the errors are normally distributed, we considered the following eight scenarios for the distribution of the carriers:

- In the first three, all coordinates of $\mathbf{x}$ are independent, with distribution $N(0, 1)$, $U(0, 1)$ and Student's t-distribution with 4 degrees of freedom.
- In the next three, we include an intercept: $(x_2, \ldots, x_p)$ are as above, and $x_1 = 1$.

Table 6.2 Minimum efficiencies of the estimators under standard normal errors, over the eight scenarios for the carriers

p	n	MM85	B-S	REWLSE	BRS	DCML	S
5	25	0.522	0.969	0.558	0.784	0.809	0.231
	50	0.754	0.994	0.801	0.951	0.948	0.250
	100	0.814	0.997	0.882	0.983	0.981	0.266
10	50	0.600	0.990	0.639	0.898	0.880	0.235
	100	0.774	0.998	0.847	0.986	0.990	0.265
	200	0.816	0.999	0.910	0.987	0.996	0.256
20	100	0.658	0.999	0.742	0.971	0.961	0.272
	200	0.796	0.999	0.903	0.995	0.994	0.262
	400	0.826	0.999	0.939	0.993	0.998	0.241

- In the last two we include quadratic terms: $(x_2, \ldots, x_p)$ are the squares of standard normal and uniform variables and $x_1 = 1$.

In all cases, the distributions were normalized so that $\mathbf{V_x} = \mathbf{I}$.

For each n, p and scenario for the carriers, we estimated the finite-sample efficiency of an estimator as the ratio between the MSE of the LSE and that of the estimator in question.

To summarize the results for the case of normal errors, for each combination (p, n), we took for each estimator the minimum finite-sample efficiency over the eight scenarios for the carriers. Table 6.2 displays these results.

We note that:

- The S-estimator is very inefficient in all cases.
- The finite sample efficiency of both the MM85 and REWLSE can be much lower than their asymptotic efficiency for small n/p. Their worst behaviour arises when the carriers are the squares of standard normal random variables.
- BRS and DCML outperform MM85 and REWLSE, and are highly efficient in all cases.
- B-S shows the highest efficiencies in all cases.

We also studied the efficiency of the estimators when the errors have Student's t-distribution with 3 and 5 degrees of freedom (d.f.) and the carriers are independent with $N(0, 1)$ distribution plus an intercept. In this case, the efficiency is estimated by the ratio between the MSE of the Student MLE and the MSE of the corresponding estimator. The results are shown in Table 6.3.

In this case all the estimators, except the S-estimator, are highly efficient. Note that the efficiency of BRS is always higher than that of DCML.

Table 6.3 Efficiencies of the estimators under errors with Student's t-distribution with 3 and 5 degrees of freedom and normal carriers

d.f.	p	n	MM85	B-S	REWLSE	BRS	DCML	S
3	5	25	0.797	0.889	0.777	0.898	0.865	0.430
		50	0.897	0.899	0.833	0.917	0.882	0.415
		100	0.944	0.901	0.873	0.903	0.877	0.455
	10	50	0.844	0.864	0.835	0.939	0.908	0.400
		100	0.942	0.905	0.878	0.889	0.863	0.413
		200	0.943	0.896	0.856	0.889	0.869	0.454
	20	100	0.887	0.916	0.889	0.929	0.897	0.387
		200	0.942	0.885	0.875	0.909	0.879	0.403
		400	0.960	0.891	0.883	0.899	0.878	0.453
5	5	25	0.735	0.933	0.759	0.902	0.895	0.373
		50	0.891	0.947	0.876	0.962	0.942	0.377
		100	0.921	0.935	0.891	0.964	0.934	0.375
	10	50	0.770	0.952	0.791	0.954	0.918	0.319
		100	0.887	0.952	0.888	0.979	0.946	0.358
		200	0.930	0.945	0.900	0.956	0.915	0.369
	20	100	0.828	0.940	0.867	0.989	0.960	0.343
		200	0.910	0.917	0.911	0.978	0.947	0.342
		400	0.936	0.935	0.919	0.979	0.937	0.366

6.4.2 Robustness

To assess the estimators' robustness we contaminated the data with a fraction $\varepsilon = 0.1$ and 0.2 of outliers as follows. Let $m = [n\varepsilon]$, then for $i \leq n - m$, $(\mathbf{x}_i^{\mathrm{T}}, y_i)$ were generated according to (6.1), where the errors u_i have standard normal distribution, $\mathbf{x} = (1, x_2, \ldots, x_p)$, that is the first carrier corresponds to an intercept, and $x_2, \ldots, x_p$ are independent with standard normal distribution. For $i > n - m$ we took $\mathbf{x}_i = (1, x_0, 0, \ldots, 0)^{\mathrm{T}}$ and $y_i = x_0 k$. The effect of this contamination will be to artificially drag the estimators to $(0, k, 0, \ldots, 0)^{\mathrm{T}}$. We took $x_0 = 5$ and moved k in an uniformly spaced grid between 0.5 and 10 with step 0.1.

Figure 6.1 displays the MSEs of the estimators for $p = 10$, $n = 50$, normal carriers and $\varepsilon = 0.1$ for different values of the outlier size k. The upper panel shows that MM85, REWLSE, DCML and BRS have similar behaviours, the maximum MSE of BRS being the lowest. The lower panel shows that the MSEs of the S-estimator and B-S are generally higher than that of BRS, and that the maximum MSE of B-S is the highest of all the estimators considered in the simulation.

Table 6.4 shows for each (p, n) the maximum mean squared error attained by each estimator.

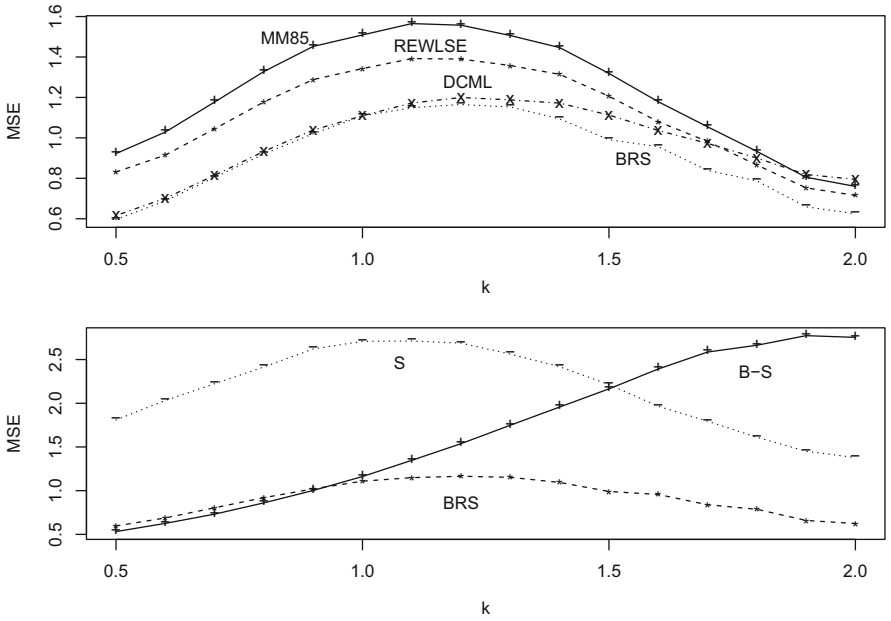

Fig. 6.1 MSEs of regression estimators as function of outlier size k for normal $\mathbf{x}$, $p = 10$, $n = 50$ and $\varepsilon = 0.1$. Note that the y-axis of the upper panel is zoomed in from the y-axis of the lower panel

Table 6.4 Maximum mean squared errors of estimators with normal predictors for contaminated data

ε	p	n	MM85	B-S	REWLSE	BRS	DCML	S
0.1	5	25	1.10	2.09	0.99	0.93	0.94	1.81
		50	0.72	1.71	0.62	0.61	0.60	1.25
		100	0.48	1.61	0.40	0.43	0.40	0.83
	10	50	1.57	2.77	1.39	1.16	1.20	2.71
		100	0.69	1.83	0.58	0.58	0.55	1.30
		200	0.46	1.69	0.39	0.42	0.39	0.84
	20	100	1.52	3.37	1.30	1.11	1.11	2.58
		200	0.69	2.28	0.56	0.56	0.55	1.36
		400	0.50	1.92	0.41	0.45	0.42	0.95
0.2	5	25	10.74	28.49	10.32	9.06	8.45	12.30
		50	3.93	11.37	3.61	3.38	3.38	5.34
		100	2.34	7.94	2.12	2.19	2.12	3.27
	10	50	11.73	26.57	11.39	9.78	9.31	13.55
		100	3.62	13.03	3.29	3.12	3.09	5.10
		200	2.13	9.67	1.90	1.98	1.92	3.03
	20	100	10.35	31.23	9.84	8.39	8.10	12.96
		200	3.94	13.45	3.52	3.39	3.36	5.08
		400	2.58	10.85	2.28	2.37	2.34	3.58

Note that for $\varepsilon = 0.1$ the maximum MSEs of REWLSE, DCML and BRS are very similar, and are in all cases lower than those of and MM85, which in turn are lower than those of S. BRS has the lowest maximum MSE for $n/p = 5$, DCML has the lowest maximum MSE for $n/p = 10$ and REWLSE has the lowest maximum MSE for $n/p = 20$. For $\varepsilon = 0.2$, DCML generally shows the best performance for $n/p = 5$ and $n/p = 10$, followed closely by BRS, whereas REWLSE has the lowest maximum MSE for $n/p = 20$.

B-S has the highest maximum MSE in all cases, showing that the price to pay for its remarkably high efficiency is an important loss in robustness.

6.4.3 Real Data

In this section we apply the estimators included in the simulation study to a well-known data set, the Hertzprung-Russell diagram of the star cluster CYG OB1. This dataset, which consists of the temperature and light intensity of 47 stars in the direction of Cygnus, was previously analysed in Rousseeuw and Leroy (1987). The carrier is the logarithm of the effective temperature at the surface of the star (T), and the response variable is the logarithm of its light intensity (L/L_0). The Hertzprung-Russell diagram is shown in Fig. 6.2. It is seen that the observations are roughly split into two groups: one containing most of the observations and the other one the observations labelled 11, 20, 30 and 34. The stars in this last group are called giants, and the rest of the stars are said to lie on the main sequence. The giant stars do not follow the same pattern as the majority of the data and for this reason they are considered outliers. Figure 6.2 also shows the BRS, B-S and LSE fits. It is seen that BRS gives a good fit for the majority of the observations, whereas B-S and LSE

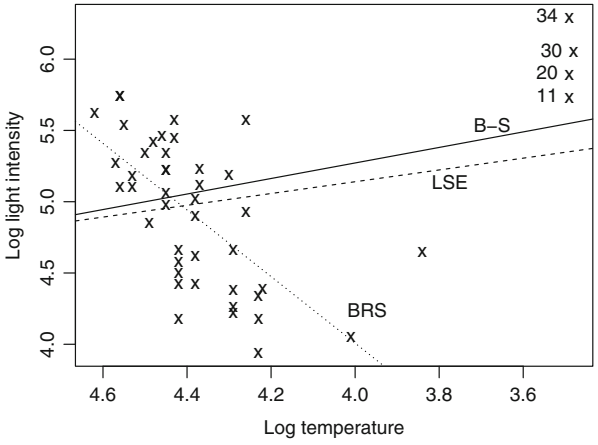

Fig. 6.2 Hertzsprung-Russell diagram with LSE, B-S and BRS fits

Table 6.5 Hertzprung-Russell diagram data: RMSEs calculated using only the residuals of stars lying in the main sequence

Computed with	LSE	MM85	B-S	REWLSE	BRS	DCML	S
Giants removed	0.3962	0.4227	0.3963	0.4268	0.3995	0.4054	0.4790
Whole data	0.5401	0.4105	0.5677	0.4127	0.3988	0.3990	0.4404

make a compromise between the two groups of observations, and thus fit neither one of them well.

Lacking a "true model", we employ an alternative criteria to evaluate the estimators' robustness and efficiency. First, we calculated each of the estimators using only the data of the stars that lie on the main sequence, i.e. without the giant stars (observations 11, 20, 30, and 34), and then calculated the root mean squared error (RMSE, the square root of the mean of the squared residuals) using the same data. We compared the RMSEs of the estimators with that of LSE as a surrogate criterion for efficiency. For a surrogate criterion for robustness, we calculated each of the estimators using all of the data, and then calculated the RMSE using only the data of the stars lying in the main sequence. Results are shown in Table 6.5.

It is seen that B-S is the most efficient of all the estimators. BRS is the second most efficient followed by DCML, MM85, REWLSE and S. Regarding the robustness of the estimators, BRS shows the best behaviour, followed by DCML, MM85, REWLSE, S, LSE and B-S in that order.

6.5 Conclusions

In this paper we proposed a new family of robust regression estimators, which are simultaneously highly robust and highly efficient for small samples with normally distributed errors. In an extensive simulation study this estimator was compared with other robust estimators which have similar properties.

A joint analysis of the results shown in Tables 6.2, 6.3 and 6.4 shows that DCML and BRS exhibit the best balance between robustness and efficiency. BRS is more efficient for errors with heavy tails like Student's t-distribution with 3 and 5 degrees of freedom and DCML is generally more robust under a contaminated normal distribution.

We should note that BRS-estimators can be extended for another models where S- and MM-estimators can be defined. For example, to estimate multivariate location and scatter (see Davies (1987) for S-estimators and Lopuhaä (1992) for MM-estimators), or for non-linear regression (see Fasano et al. (2012) for S- and MM-estimators). However these extensions are a matter of further research.

An R code to compute the recommended BRS-estimator is available at http://mate.dm.uba.ar/~vyohai/BRSreg.r.

Acknowledgements This research was partially supported by Grants W276 from Universidad of Buenos Aires, PIPs 112-2008-01-00216 and 112-2011-01- 00339 from CONICET and PICT 2011-0397 from ANPCYT, Argentina. Ezequiel Smucler was supported by a doctoral scholarship of the University of Buenos Aires. We are grateful to two anonymous referees, whose comments lead to improvements in the paper.

References

Bondell, H.D., Stefanski, L.A.: Efficient robust regression via two-stage generalized empirical likelihood. J. Am. Stat. Assoc. **108**, 644–655 (2013)

Davies, P.L.: Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. Ann. Stat. **15**, 1269–1292 (1987)

Davies, L.: The asymptotics of S-estimators in the linear regression model. Ann. Stat. **18**, 1651–1675 (1990)

Donoho, D.L., Huber, P.J.: The notion of breakdown point. In: Festschrift for Erich L. Lehmann, pp. 157–184. Wadsworth, Belmont (1983)

Fasano M.V., Maronna R.A., Sued R.M., Yohai V.J.: Continuity and differentiability of regression M functionals. Bernoulli **18**, 1289–1309 (2012)

Gervini, D., Yohai, V.J.: A class of robust and fully efficient regression estimators. Ann. Stat. **30**, 583–616 (2002)

Hampel, F.R.: Beyond location parameters: robust concepts and method (with discussion). Bull. Int. Stat. Inst. **46**, 375–391 (1975)

Hössjer, O.: On the optimality of S-estimators. Stat. Probab. Lett. **14**, 413–419 (1992)

Huber, P.J.: Robust Statistics. Wiley, New York (1981)

Lopuhaä H.P.: Highly efficient estimators of multivariate location with high breakdown point. Ann. Stat. **20**, 398–413 (1992)

Maronna, R.A., Yohai, V.J.: High finite-sample efficiency and robustness based on distance-constrained maximum likelihood. Comput. Stat. Data Anal. (2014). doi:10.1016/j.csda.2014.10.015

Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust Statistics: Theory and Methods. Wiley, New York (2006)

Rousseeuw, P.J.: Least median of squares regression. J. Am. Stat. Assoc. **79**, 871–880 (1984)

Rousseeuw, P.J., Leroy, A.: Robust Regression and Outlier Detection. Wiley, New York (1987)

Rousseeuw, P.J., Yohai, V.J.: Robust Regression by Means of S-Estimators. Lecture Notes in Statistics, vol. 26, pp. 256–272. Springer, New York (1984)

Stahel, W.A., Yohai, V.J., Zamar, R.H.: A procedure for robust estimation and inference in linear regression. In: Stahel, W., Weisberg, S. (eds) Directions in Robust Statistics and Diagnostics (Part II). The IMA Volumes in Mathematics and Its Applications. Springer, New York (1991)

Yohai, V.J.: High Breakdown point and high efficiency estimates for regression. Ann. Stat. **15**, 642–656 (1987)

Yohai, V.J., Zamar, R.H.: High breakdown point estimates of regression by means of an efficient scale. Technical Report No. 84, Department of Statistics, University of Washington, Seattle. Available at https://www.stat.washington.edu/research/reports/1986/tr084.pdf (1986)

Chapter 7
Optimal Rank Tests for Symmetry Against Edgeworth-Type Alternatives

Delphine Cassart, Marc Hallin, and Davy Paindaveine

Abstract We are constructing, for the problem of univariate symmetry (with respect to specified or *un*specified location), a class of signed-rank tests achieving optimality against the family of asymmetric (local) alternatives considered in Cassart et al. (Bernoulli 17:1063–1094, 2011). Those alternatives are based on a non-Gaussian generalization of classical first-order Edgeworth expansions indexed by a measure of skewness such that *(1)* location, scale, and skewness play well-separated roles (diagonality of the corresponding information matrices), and *(2)* the classical tests based on the Pearson–Fisher coefficient of skewness are optimal in the vicinity of Gaussian densities. Asymptotic distributions are derived under the null and under local alternatives. Asymptotic relative efficiencies are computed and, in most cases, indicate that the proposed rank tests significantly outperform their traditional competitors.

Keywords Asymptotic relative efficiency • Edgeworth expansion • Local asymptotic normality • Locally asymptotically optimal test • Signed rank test • Skewness • Symmetry

D. Cassart
ECARES, Université Libre de Bruxelles, 1050 Bruxelles, Belgium
e-mail: dcassart@gmail.com

M. Hallin (✉)
ECARES, Université Libre de Bruxelles, 1050 Bruxelles, Belgium
ORFE, Princeton University, Princeton, NJ 08544, USA
e-mail: mhallin@ulb.ac.be

D. Paindaveine
ECARES and Department of Mathematics, Université Libre de Bruxelles, 1050 Bruxelles, Belgium
e-mail: dpaindav@ulb.ac.be

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_7

7.1 Introduction

The assumption of symmetry is among the most important and fundamental ones in statistics. This importance explains the variety of existing parametric as well as nonparametric testing procedures of the null hypothesis of symmetry in an i.i.d. sample $X_1, \ldots, X_n$; see Hollander (1988) for a classical survey.

Traditional parametric tests of the null hypothesis of symmetry—the hypothesis under which $X_1 - \theta \stackrel{d}{=} -(X_1 - \theta)$ for some location (automatically, the population median) $\theta \in \mathbb{R}$, where $\stackrel{d}{=}$ stands for equality in distribution—are based on third-order moments. Write $m_k^{(n)} := m_k^{(n)}(\bar{X}^{(n)})$ for the sample moment of order k, where

$$m_k^{(n)}(\theta) := n^{-1} \sum_{i=1}^{n} (X_i - \theta)^k$$

and $\bar{X}^{(n)} := n^{-1} \sum_{i=1}^{n} X_i$. When the location θ is specified, the classical test statistic is

$$S_1^{(n)} := n^{1/2} m_3^{(n)}(\theta) / (m_6^{(n)}(\theta))^{1/2}, \tag{7.1}$$

with, under finite sixth-order moments, asymptotically standard normal null distribution. When θ is unspecified, the classical test is based on the empirical coefficient of skewness $b_1^{(n)} := m_3^{(n)} / s_n^3$, where $s_n := (m_2^{(n)})^{1/2}$ stands for the empirical standard error in a sample of size n. More precisely, that test relies on the asymptotic standard normal distribution (still under finite moments of order six) of

$$S_2^{(n)} := n^{1/2} m_3^{(n)} / (m_6^{(n)} - 6 s_n^2 m_4^{(n)} + 9 s_n^6)^{1/2}. \tag{7.2}$$

This test is generally considered a Gaussian test. Cassart et al. (2011) indeed show that it is locally asymptotically optimal, in the Le Cam sense, for the null hypothesis of i.i.d. Gaussian observations with unspecified location and scale, against asymmetric alternatives described by a first-order Edgeworth expansion of the form

$$\phi(x - \theta) + n^{-1/2} \xi (x - \theta) \phi(x - \theta)((x - \theta)^2 - \kappa), \tag{7.3}$$

where $\kappa (= 3)$ is the Gaussian kurtosis coefficient, θ a location parameter, and ξ a measure of skewness; see Chap. XVI of Feller (1971) for a concise introduction to those expansions, the idea of which goes back to Edgeworth (1905). Cassart et al. (2011) further show that the local experiments associated with (7.3) enjoy the appealing property that location, scale, and skewness play well-separated roles, in the sense that the Fisher information matrix associated with the triple $(\theta, \sigma\ \xi)$ is diagonal.

Extending that Gaussian approach to a broad class of symmetric densities by embedding the null hypothesis of i.i.d.-ness with density

$$f: \ x \mapsto f(x) := \sigma^{-1} f_1((x-\theta)/\sigma)$$

(f_1 symmetric with respect to the origin) into a family of locally asymmetric alternatives based on an adequate generalization of (7.3), Cassart et al. (2011) show that those families (for fixed f) are uniformly locally asymptotically normal (ULAN) under mild assumptions on f_1. The resulting locally asymptotically optimal tests are derived and studied in detail.

It should be insisted, however, that considering alternatives of this (generalized) Edgeworth type by no means implies any assumption on the form of the asymmetry present in the data. Those local expansions are just considered as a way of producing non-Gaussian alternatives to the classical test (7.2). And simulations (see Cassart et al. 2011) indicate that the resulting tests perform quite well against other types of asymmetries, such as the skew-normal and skew-t densities investigated, for instance, by Azzalini and Capitano (2003).

Yet, these tests all are of a parametric nature, while symmetry, very typically, is a nonparametric hypothesis, enjoying a rich group invariance structure, through which classical maximal invariance arguments naturally bring *signs* and *signed-ranks* into the picture.

The most popular nonparametric signed-rank tests of symmetry (with respect to any specified location—without loss of generality, the origin) are the *sign test*, based on the binomial null distribution of the number of negative signs in a sample of size n, and the *Wilcoxon signed-rank test*, based on the exact or asymptotic null distribution of

$$S_W^{(n)} := n^{-1/2} \sum_{i=1}^{n} s_i R_{+,i}^{(n)}$$

(or a linear transformation thereof, such as $n^{-1/2} \sum_{i=1}^{n} I[s_i = 1] R_{+,i}^{(n)}$), where $s_1, \ldots, s_n$ denote the signs, and $R_{+,1}^{(n)}, \ldots, R_{+,n}^{(n)}$ the ranks of absolute values in a sample of size n. These tests are not optimal in any satisfactory sense against asymmetry: actually, they are locally asymptotically most powerful against symmetry-preserving location shifts—under otherwise unspecified density for the sign test, under logistic densities for the Wilcoxon one. Moreover, the sign test is completely insensitive to nonsymmetric alternatives preserving the median. There is no way such tests can be adapted to an unspecified-location context (*alignment* for the location nuisance here produces, at probability level α, a test with trivial asymptotic power α).

Another signed-rank test based on signs only is the runs test proposed by McWilliams (1990) and further investigated by Henze (1993). If not completely invariant, this test has low sensitivity against location shifts; however, it does not address any well-identified alternative and does not exploit the ranks themselves.

The *triples test* by Randles et al. (1980) is location-invariant, and also based on signs. Those signs, though, are those of quantities of the form $X_i + X_j - 2X_k$, $1 \leq i < j \leq n$ and $i \neq j \neq k$, which do not follow from any concept of group invariance and are not distribution-free; optimality properties, if any, are unclear.

To the best of our knowledge, the problem of constructing optimal rank-based tests of symmetry only has been touched in Cassart et al. (2008) and Ley and Paindaveine (2009), both focusing on quite specific alternatives (Fechner and Ferreira-Steel types, respectively). The objective of this paper is to construct signed-rank tests that are optimal against the Edgeworth-type alternatives defined in Cassart et al. (2011). The proposed tests are distribution-free (asymptotically so in case of an unspecified location θ) under the null hypothesis of symmetry, and therefore remain valid under very mild distributional assumptions (for the specified-location case, they are valid in the absence of *any* distributional assumption).

For instance, the normal-score signed-rank test rejects the null hypothesis of symmetry with respect to (specified) θ for large values of

$$\mathcal{T}^{(n)}_{\phi_1}(\theta) := \left(n\mathcal{Y}^{(n)}_{\phi_1}\right)^{-1/2} \sum_{i=1}^{n} s_i(\theta)\Phi^{-1}\Big(\frac{n+1+R^{(n)}_{+,i}(\theta)}{2(n+1)}\Big) \times\Big(\Big(\Phi^{-1}\Big(\frac{n+1+R^{(n)}_{+,i}(\theta)}{2(n+1)}\Big)\Big)^2 - 3\Big),$$

where Φ denotes the standard normal distribution function, $s_i(\theta)$ is the sign of $Z_i(\theta) := X_i - \theta$, $R^{(n)}_{+,i}(\theta)$ the rank of $|Z_i(\theta)|$ among $|Z_1(\theta)|, \ldots, |Z_n(\theta)|$, and

$$\mathcal{Y}^{(n)}_{\phi_1} := n^{-1} \sum_{r=1}^{n} \left(\Phi^{-1}\Big(\frac{n+1+r}{2(n+1)}\Big)\Big(\Big(\Phi^{-1}\Big(\frac{n+1+r}{2(n+1)}\Big)\Big)^2 - 3\Big)\right)^2$$

a standardizing constant. That test is distribution-free under the null hypothesis of symmetry with respect to θ, asymptotically equivalent to the test (7.2) based on $b^{(n)}_1$ under Gaussian densities, and hence asymptotically most powerful against local Edgeworth alternatives of the form (7.3) with $\xi > 0$. And, under a very broad class of non-Gaussian densities (containing, among many others, all Student and power-exponential ones), the ARE (see Sect. 7.3.4) of this signed-rank test is strictly larger than one with respect to the traditional test based on $b^{(n)}_1$.

The problem we are considering throughout is that of testing the null hypothesis of symmetry. In the notation of Sect. 7.1, ξ (see (7.4) for a more precise definition) is thus the parameter of interest; the location θ and the standardized null symmetric density f_1 either are specified or play the role of nuisance parameters, whereas the scale σ (not necessarily a standard error) always is a nuisance.

The paper is organized as follows. In Sect. 7.2.1, we briefly describe the Edgeworth-type families of local alternatives we are considering. Section 7.2.2 restates the local asymptotic normality (with respect to location, scale, and asymmetry parameters) result of Cassart et al. (2011) that provides the main theoretical tool of the paper. Signed-rank versions of the corresponding central sequences for asymmetry are defined in Sect. 7.2.3. In Sect. 7.3.1, we propose nonparametric signed-rank (hence distribution-free) versions of the optimal procedures obtained in Cassart et al. (2011) for specified location θ. The case of unspecified θ is treated in Sect. 7.3.2, and requires the delicate estimation of a cross-information quantity of the same type as those appearing in the asymptotic variances of R-estimators (see Cassart et al. 2010). That estimation problem is discussed in some detail in Sect. 7.4. The van der Waerden, Wilcoxon, and Laplace versions of the signed-rank statistics are described in Sect. 7.3.3, and Sect. 7.3.4 provides asymptotic relative efficiencies of signed-rank tests with respect to the classical ones based on (7.1) and (7.2), indicating the superiority of the former.

7.2 A Class of Locally Asymptotically Normal Families of Asymmetric Distributions

7.2.1 Families of Asymmetric Densities Based on Edgeworth Approximations

Denote by $\boldsymbol{X}^{(n)} := (X_1^{(n)}, \ldots, X_n^{(n)})$, $n \in \mathbb{N}$ an i.i.d. n-tuple of observations with common density f. The null hypotheses we are interested in are

(a) the hypothesis $\mathscr{H}_\theta^{(n)}$ of symmetry with respect to specified location $\theta \in \mathbb{R}$: under $\mathscr{H}_\theta^{(n)}$, the X_i's have density function $x \mapsto f(x) := \sigma^{-1} f_1((x-\theta)/\sigma)$ (all densities are over the real line, with respect to the Lebesgue measure), for some unspecified $\sigma \in \mathbb{R}_0^+$ and f_1 in the class of standardized symmetric densities

$$\mathscr{F}_0 := \left\{ f_1 \ : \ f_1(-z) = f_1(z) \text{ and } \int_{-1}^{1} f_1(z)\, dz = 0.5 \right\}.$$

The scale parameter σ (associated with the symmetric density f) we are considering here thus is not the standard error, but the median of the absolute deviations $|X_i - \theta|$; this avoids making any moment assumptions on f;

(b) the hypothesis $\mathscr{H}^{(n)} := \bigcup_{\theta \in \mathbb{R}} \mathscr{H}_\theta^{(n)}$ of symmetry with respect to unspecified location.

As explained in the introduction, efficient testing requires the definition of families of asymmetric alternatives exhibiting some adequate structure, such as local asymptotic normality, at the null hypothesis of symmetry. For a selected class of densities f enjoying the required regularity assumptions, we therefore are

embedding the null hypothesis of symmetry into families of distributions indexed by $\theta \in \mathbb{R}$ (location), $\sigma \in \mathbb{R}_0^+$ (scale), and a parameter $\xi \in \mathbb{R}$ characterizing asymmetry. More precisely, consider the class $\mathscr{F}_1$ of densities f_1 satisfying

(i) (symmetry and standardization) $f_1 \in \mathscr{F}_0$;
(ii) (absolute continuity) there exists $\dot{f}_1$ such that, for all finite $z_1 < z_2$,

$$f_1(z_2) - f_1(z_1) = \int_{z_1}^{z_2} \dot{f}_1(z)\, dz;$$

(iii) (strong unimodality) $z \mapsto \phi_{f_1}(z) := -\dot{f}_1(z)/f_1(z)$ is monotone increasing, or the difference of two monotone increasing functions, and
(iv) (finite Fisher information) the integral

$$\mathscr{K}(f_1) := \int_{-\infty}^{+\infty} z^4 \phi_{f_1}^2(z) f_1(z) dz$$

is finite, hence also, under (iii) above, the integrals

$$\mathscr{I}(f_1) := \int_{-\infty}^{+\infty} \phi_{f_1}^2(z) f_1(z) dz \quad \text{and} \quad \mathscr{J}(f_1) := \int_{-\infty}^{+\infty} z^2 \phi_{f_1}^2(z) f_1(z) dz;$$

(v) there exists $\beta > 0$ such that

$$\int_a^{\infty} f_1(z)\, dz = O(a^{-\beta}) \quad \text{as } a \to \infty \quad \text{and} \quad \phi_{f_1}(z) = o(z^{\beta/2-2}) \quad \text{as } z \to \infty.$$

That class $\mathscr{F}_1$ thus consists of all symmetric standardized densities f_1 that are absolutely continuous, strongly unimodal (that is, log-concave), and have finite information $\mathscr{I}(f_1)$ for location, $\mathscr{J}(f_1)$ for scale and, as we shall see, $\mathscr{K}(f_1)$ for asymmetry, with tails satisfying (v). Assumption (iii) is not required for ULAN, but for the asymptotic representation of the rank-based statistics associated with the score function ϕ_{f_1}.

For all $f_1 \in \mathscr{F}_1$, denote by $\kappa(f_1) := \mathscr{J}(f_1)/\mathscr{I}(f_1)$ the ratio of information for scale and information for location. For Gaussian densities ($f_1 = \phi_1$), $\kappa(\phi_1) = 3$ reduces to kurtosis and, as we shall see, $\kappa(f_1)$ can be interpreted as a *generalized kurtosis coefficient*. Finally, write $\mathrm{P}^{(n)}_{\theta,\sigma,\xi;f_1}$ for the probability distribution of $\mathbf{X}^{(n)}$ when the X_i's are i.i.d. with density

$$\begin{aligned} f(x) = {} & \sigma^{-1} f_1\left(\frac{x-\theta}{\sigma}\right) \\ & - \frac{\xi}{\sigma}\dot{f}_1\left(\frac{x-\theta}{\sigma}\right)\left(\left(\frac{x-\theta}{\sigma}\right)^2 - \kappa(f_1)\right) I[|x-\theta| \le \sigma|z^*|] \\ & + \frac{\mathrm{sign}(\xi)}{\sigma} f_1\left(\frac{x-\theta}{\sigma}\right)\Big[I[x-\theta > \mathrm{sign}(-\xi)\sigma|z^*|] - I[x-\theta < \mathrm{sign}(\xi)\sigma|z^*|]\Big]. \end{aligned} \tag{7.4}$$

Here $\theta \in \mathbb{R}$ and $\sigma \in \mathbb{R}_0^+$ clearly are location and scale parameters, $\xi \in \mathbb{R}$ is a measure of skewness, $\kappa(f_1)$ (strictly positive for $f_1 \in \mathscr{F}_1$) the generalized kurtosis coefficient just defined, and z^* the unique (for ξ small enough; unicity follows from the monotonicity of $\phi_{f_1}(z)$) solution of $f_1(z^*) = \xi \dot{f}_1(z^*)((z^*)^2 - \kappa(f_1))$. The function f defined in (7.4) is indeed a probability density (nonnegative, integrating up to one), since it is obtained by adding and subtracting the same probability mass

$$\frac{|\xi|}{\sigma} \int_\theta^\infty \min\left(\dot{f}_1\left(\frac{x-\theta}{\sigma}\right)\left(\left(\frac{x-\theta}{\sigma}\right)^2 - \kappa(f_1)\right), f_1\left(\frac{x-\theta}{\sigma}\right)\right) dx$$

on both sides of θ (according to the sign of ξ). Note that $\xi > 0$ implies $f(x) = 0$ for $x - \theta < -\sigma|z^*|$ and $f(x) = 2\sigma^{-1} f_1((x-\theta)/\sigma)$ for $x - \theta > \sigma|z^*|$. Moreover, the density $x \mapsto f(x)$ is continuous whenever $\dot{f}_1(x)$ is; it vanishes for $x \le \theta + \sigma z^*$ if $\xi > 0$, for $x \ge \theta + \sigma z^*$ if $\xi < 0$, and is left- or right-skewed according as $\xi < 0$ or $\xi > 0$. As for z^*, it tends to $-\infty$ as $\xi \downarrow 0$, to ∞ as $\xi \uparrow 0$; in the Gaussian case, it is easy to check that $|z^*| = O(|\xi|^{-1/3})$ as $\xi \to 0$.

The intuition behind this class of alternatives is that, in the Gaussian case, Eq. (7.4), with $\xi = n^{-1/2}\tau$ yields (for $x \in [\theta \pm \sigma z^*]$) the first-order Edgeworth expansion of the density of the standardized mean of an i.i.d. n-tuple of variables with third-order moment $6\tau\sigma^3$ (where standardization is based on the median σ of absolute deviations from θ). For a "local" value of ξ, of the form $n^{-1/2}\tau$, (7.4) thus describes the type of deviation from symmetry that corresponds to the classical central-limit context. Hence, if a Gaussian density is justified as resulting from the additive combination of a large number of small independent symmetric variables, the locally asymmetric f results from the same additive combination, of independent, but slightly skewed ones. We show in Cassart et al. (2011) that the locally optimal test in such case is the traditional test based on $b_1^{(n)}$.

Besides the Gaussian one, interesting special cases of (7.4) are obtained in the vicinity of

(i) the Student distributions with $\nu > 2$ degrees of freedom, with standardized densities

$$f_1(z) = f_{t_\nu}(z) := C_{t_\nu}(1 + a_\nu z^2/\nu)^{-(\nu+1)/2},$$
$$\mathscr{I}(f_1) = a_\nu(\nu+1)/(\nu+3), \quad \mathscr{J}(f_1) = 3(\nu+1)/(\nu+3),$$

and

$$\mathscr{K}(f_1) = 15\nu(\nu+1)/a_\nu(\nu-2)(\nu+3);$$

the corresponding Gaussian values (density $\phi_1(z) := (a/2\pi)^{1/2}\exp(-az^2/2)$) are obtained by taking limits as $\nu \to \infty$: $\mathscr{I}(\phi_1) = a \approx 0.4549$, $\mathscr{J}(\phi_1) = 3$ and $\mathscr{K}(\phi_1) = 15/a$;

(ii) the logistic distributions, with standardized density

$$f_1(z) = f_{\text{Log}}(z) := \sqrt{b}\exp(-\sqrt{b}z)/(1+\exp(-\sqrt{b}z))^2,$$
$$\mathscr{I}(f_1) = b/3, \quad \mathscr{J}(f_1) = (12+\pi^2)/9, \quad \text{and} \quad \mathscr{K}(f_1) = \pi^2(120+7\pi^2)/45b;$$

(iii) the double-exponential (or Laplace) distributions, with standardized density

$$f_1(z) = f_{\mathscr{L}}(z) := (1/2d)\exp(-|z|/d),$$
$$\mathscr{I}(f_1) = 1/d^2, \quad \mathscr{J}(f_1) = 2, \quad \text{and} \quad \mathscr{K}(f_1) = 24d^2;$$

(iv) the power-exponential distributions, with standardized densities

$$f_1(z) = f_{\text{exp}_\eta}(z) := C_{\text{exp}_\eta}\exp(-(g_\eta z)^{2\eta}),\ \eta \in \mathbb{N}_0,$$
$$\mathscr{I}(f_1) = 2g_\eta^2\eta\frac{\Gamma(2-1/2\eta)}{\Gamma(1+1/2\eta)}, \quad \mathscr{J}(f_1) = 1+2\eta, \text{ and } \mathscr{K}(f_1) = 2g_\eta\frac{\eta}{\Gamma(1+1/2\eta)}$$

(the positive constants C_{t_ν}, C_{exp_η}, a_ν, a, b, d, and g_η are such that $f_1 \in \mathscr{F}_1$). Figure 7.1 provides graphical representations of some densities in the Gaussian ($f_1 = \phi_1$) and double-exponential ($f_1 = f_{\mathscr{L}}$) Edgeworth families (7.4), respectively. In the Gaussian case, the skewed densities are continuous, while the double-exponential case, due to the discontinuity of $\dot{f}_{\mathscr{L}}(x)$ at $x = 0$, exhibits a discontinuity at the origin.

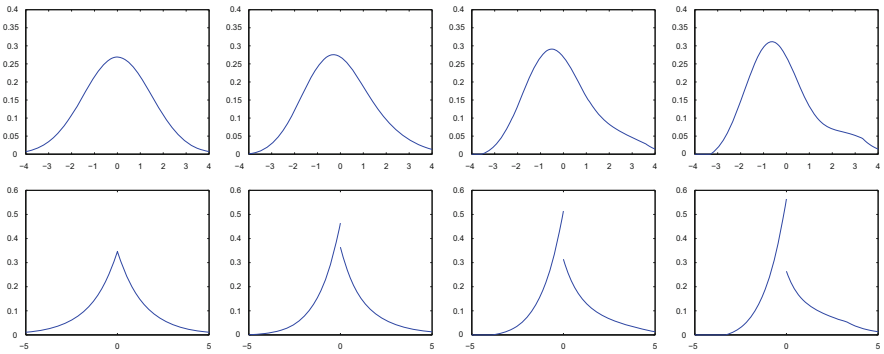

Fig. 7.1 Graphical representations of the Gaussian ($f_1 = \phi_1$) and double-exponential ($f_1 = f_{\mathscr{L}}$) Edgeworth families (7.4), for asymmetry parameter values $\xi = 0,\ 0.05,\ 0.10$, and 0.15 (from left to right)

7.2.2 *Uniform Local Asymptotic Normality*

The main technical tool in our derivation of optimal tests is the uniform local asymptotic normality (ULAN), with respect to $\boldsymbol{\vartheta} := (\theta, \sigma, \xi)'$, at any $(\theta, \sigma, 0)'$, of the parametric families

$$\mathscr{P}^{(n)}_{f_1} := \bigcup_{\sigma>0} \mathscr{P}^{(n)}_{\sigma;f_1} := \bigcup_{\sigma>0} \left\{ \mathrm{P}^{(n)}_{\theta,\sigma,\xi;f_1} \mid \theta \in \mathbb{R},\ \xi \in \mathbb{R} \right\}, \tag{7.5}$$

where $f_1 \in \mathscr{F}_1$. More precisely, we are using the following result, which is proved in Cassart et al. (2011).

Proposition 7.1 (ULAN) *For any $f_1 \in \mathscr{F}_1$, $\theta \in \mathbb{R}$, and $\sigma \in \mathbb{R}^+_0$, the family $\mathscr{P}^{(n)}_{f_1}$ is ULAN at $(\theta, \sigma, 0)'$, with (writing Z_i for $Z^{(n)}_i(\theta, \sigma) := \sigma^{-1}(X^{(n)}_i - \theta)$) central sequence*

$$\boldsymbol{\Delta}^{(n)}_{f_1}(\boldsymbol{\vartheta}) =: \begin{pmatrix} \Delta^{(n)}_{f_1;1}(\boldsymbol{\vartheta}) \\ \Delta^{(n)}_{f_1;2}(\boldsymbol{\vartheta}) \\ \Delta^{(n)}_{f_1;3}(\boldsymbol{\vartheta}) \end{pmatrix} = n^{-1/2} \sum_{i=1}^{n} \begin{pmatrix} \sigma^{-1}\phi_{f_1}(Z_i) \\ \sigma^{-1}(\phi_{f_1}(Z_i)Z_i - 1) \\ \phi_{f_1}(Z_i)\left(Z_i^2 - \kappa(f_1)\right) \end{pmatrix} \tag{7.6}$$

and full-rank information matrix

$$\boldsymbol{\Gamma}_{f_1}(\boldsymbol{\vartheta}) = \begin{pmatrix} \sigma^{-2}\mathscr{I}(f_1) & 0 & 0 \\ 0 & \sigma^{-2}(\mathscr{J}(f_1) - 1) & 0 \\ 0 & 0 & \gamma(f_1) \end{pmatrix}, \tag{7.7}$$

where $\gamma(f_1) := \mathscr{K}(f_1) - \mathscr{J}^2(f_1)/\mathscr{I}(f_1)$. In other words, for any sequence $\boldsymbol{\vartheta}^{(n)}$ of the form $(\theta^{(n)}, \sigma^{(n)}, 0)'$ such that $\boldsymbol{\vartheta}^{(n)} - \boldsymbol{\vartheta} = O(n^{-1/2})$ as $n \to \infty$ for some $\boldsymbol{\vartheta} = (\theta, \sigma, 0)'$, one has that, as $n \to \infty$ under $\mathrm{P}^{(n)}_{\theta^{(n)},\sigma^{(n)},0;f_1}$,

$$\log \frac{d\mathrm{P}^{(n)}_{\theta^{(n)}+n^{-1/2}\tau^{(n)}_1,\sigma^{(n)}+n^{-1/2}\tau^{(n)}_2,n^{-1/2}+\tau^{(n)}_3;f_1}}{d\mathrm{P}^{(n)}_{\theta^{(n)},\sigma^{(n)},0;f_1}} = \boldsymbol{\tau}^{(n)\prime}\boldsymbol{\Delta}^{(n)}_{f_1}(\boldsymbol{\vartheta}^{(n)}) - \frac{1}{2}\boldsymbol{\tau}^{(n)\prime}\boldsymbol{\Gamma}_{f_1}(\boldsymbol{\vartheta})\boldsymbol{\tau}^{(n)} + o_{\mathrm{P}}(1)$$

for any bounded sequence $\boldsymbol{\tau}^{(n)} = (\tau^{(n)}_1, \tau^{(n)}_2, \tau^{(n)}_3)' \in \mathbb{R}^3$, and $\boldsymbol{\Delta}^{(n)}_{f_1}(\boldsymbol{\vartheta}^{(n)})$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Gamma}_{f_1}(\boldsymbol{\vartheta})$.

The diagonal form of the information matrix $\boldsymbol{\Gamma}_{f_1}(\boldsymbol{\vartheta})$ confirms that location, scale, and skewness, in the parametric family (7.5), play distinct and well-separated roles. Note that orthogonality between the scale and skewness components of $\boldsymbol{\Delta}^{(n)}_{f_1}(\boldsymbol{\vartheta})$ automatically follows from the symmetry of f_1, while for location and skewness,

this orthogonality is a consequence of the definition of $\kappa(f_1)$. The Gaussian versions of (7.6) and (7.7) are

$$\boldsymbol{\Delta}^{(n)}_{\phi_1}(\boldsymbol{\vartheta}) = n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} a\sigma^{-1}Z_i \\ \sigma^{-1}(aZ_i^2-1) \\ aZ_i(Z_i^2-\frac{3}{a}) \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Gamma}_{\phi_1}(\boldsymbol{\vartheta}) = \begin{pmatrix} a\sigma^{-2} & 0 & 0 \\ 0 & 2\sigma^{-2} & 0 \\ 0 & 0 & 6/a \end{pmatrix},$$

respectively (recall that $a \approx 0.4549$).

7.2.3 *Signed-Rank Versions of the Central Sequence*

As mentioned in the introduction, the hypothesis of symmetry enjoys strong group invariance features. The null hypothesis $\mathscr{H}^{(n)}_\theta$ of symmetry with respect to θ indeed is generated by the group $\mathscr{G}^{(n)}_\theta$, $\circ$ of all transformations $\mathscr{g}_h$ of $\mathbb{R}^n$ such that $\mathscr{g}_h(x_1,\ldots,x_n) := (h(x_1),\ldots,h(x_n))$, where $\lim_{x\to\infty} h(x) = \infty$, and $x \mapsto h(x)$ is continuous, monotone increasing, and skew-symmetric with respect to θ (i.e., such that $h(\theta - z) = -h(\theta + z)$). A maximal invariant for that group is known to be the vector $(s_1(\theta),\ldots,s_n(\theta))$ of signs along with the vector $(R^{(n)}_{+,1}(\theta),\ldots,R^{(n)}_{+,n}(\theta))$ of ranks, where $s_i(\theta)$ is the sign of $(X_i - \theta)$ and $R^{(n)}_{+,i}(\theta)$ the rank of $|X_i - \theta|$ among $|X_1 - \theta|,\ldots,|X_n - \theta|$.

General results on semiparametric efficiency (Hallin and Werker 2003) indicate that, in such context, the expectation of the central sequence $\boldsymbol{\Delta}^{(n)}_{f_1}(\boldsymbol{\vartheta})$ conditional on those signs and ranks yields a version of the semiparametrically efficient (at f_1 and $\boldsymbol{\vartheta}$) central sequence. The only component of the central sequence $\boldsymbol{\Delta}^{(n)}_{f_1}$ which is used in this section is the ξ-component $\Delta^{(n)}_{f_1;3}$, with *signed-rank version* (a terminology justified in Proposition 7.2)

$$\underset{\sim}{\Delta}^{(n)}_{f_1;3}(\theta) := n^{-1/2}\sum_{i=1}^{n} s_i(\theta)\phi_{f_1}\Big(F^{-1}_{1+}\Big(\frac{R^{(n)}_{+,i}(\theta)}{n+1}\Big)\Big)\Big(\Big(F^{-1}_{1+}\Big(\frac{R^{(n)}_{+,i}(\theta)}{n+1}\Big)\Big)^2 - \kappa(f_1)\Big),$$

where $F_{1+} : (x) \mapsto (2F_1(x) - 1)I[x \geq 0]$ and F_1 denote the distribution functions of $|Z_i|$ and Z_i, respectively, when Z_i has density f_1. Later on, however, we also will need the signed-rank version

$$\underset{\sim}{\Delta}^{(n)}_{f_1;1}(\theta) := n^{-1/2}\sum_{i=1}^{n} s_i(\theta)\phi_{f_1}\Big(F^{-1}_{1+}\Big(\frac{R^{(n)}_{+,i}(\theta)}{n+1}\Big)\Big)$$

of the θ-component of the same central sequence. The following results follow from the classical Hájek theory for linear signed-rank statistics (see, e.g., Chapter 3 of Puri and Sen 1985).

Proposition 7.2 *Let $f_1 \in \mathscr{F}_1$ and $g_1 \in \mathscr{F}_0$. Then,*

(i) under $\mathrm{P}^{(n)}_{\theta,\sigma,0;g_1}$, *as* $n \to \infty$,

$$\underset{\sim}{\Delta}^{(n)}_{f_1;3}(\theta) = n^{-1/2} \sum_{i=1}^{n} \phi_{f_1}\Big(F_1^{-1}(G_1(Z_i^{(n)}(\theta,\sigma)))\Big) \times\Big(\Big(F_1^{-1}(G_1(Z_i^{(n)}(\theta,\sigma)))\Big)^2 - \kappa(f_1)\Big) + o_{L^2}(1),$$

and hence, under $\mathrm{P}^{(n)}_{\theta,\sigma,0;f_1}$, $\underset{\sim}{\Delta}^{(n)}_{f_1;3}(\theta) = \Delta^{(n)}_{f_1;3}(\theta,\sigma,0) + o_{\mathrm{P}}(1)$;

(ii) under $\mathrm{P}^{(n)}_{\theta,\sigma,0;g_1}$, $\underset{\sim}{\Delta}^{(n)}_{f_1;3}(\theta)$ *has mean zero and variance*

$$\underset{\sim}{\gamma}^{(n)}_{f_1} := n^{-1} \sum_{r=1}^{n} \phi^2_{f_1}\Big(F^{-1}_{1+}\Big(\frac{r}{n+1}\Big)\Big)\Big(\Big(F^{-1}_{1+}\Big(\frac{r}{n+1}\Big)\Big)^2 - \kappa(f_1)\Big)^2 \tag{7.8}$$

tending to $\gamma(f_1)$ *as* $n \to \infty$.

Note that Part (i) of this result entails that $\underset{\sim}{\Delta}^{(n)}_{f_1;3}(\theta)$, under $\mathrm{P}^{(n)}_{\theta,\sigma,0;g_1}$, with $g_1 \in \mathscr{F}_0$, is asymptotically equivalent (in the mean square sense) to a random variable of the form $n^{-1/2}\sum_{i=1} \eta_i$, where the η_i's are i.i.d. with mean zero (since $u \mapsto \phi_{f_1}(u)u^2$ and $u \mapsto \phi_{f_1}(u)$ are skew-symmetric) and variance $\gamma(f_1)$. Consequently, $\underset{\sim}{\Delta}^{(n)}_{f_1;3}(\theta)$, still under $\mathrm{P}^{(n)}_{\theta,\sigma,0;g_1}$, is also asymptotically normal with mean zero and variance $\gamma(f_1)$.

7.3 Rank-Based Tests for Symmetry

7.3.1 Optimal Signed-Rank Tests of Symmetry: Specified Location

Proposition 7.2 immediately yields a distribution-free signed-rank test of the hypothesis of symmetry with respect to a specified location θ. With the notation of Theorem 7.2, consider the rank-based test statistic

$$\underset{\sim}{T}^{(n)}_{f_1}(\theta) := \Big(\underset{\sim}{\gamma}^{(n)}_{f_1}\Big)^{-1/2} \underset{\sim}{\Delta}^{(n)}_{f_1;3}(\theta) \tag{7.9}$$

$$= \Big(n\underset{\sim}{\gamma}^{(n)}_{f_1}\Big)^{-1/2} \sum_{i=1}^{n} s_i(\theta)\phi_{f_1}\Big(F^{-1}_{1+}\Big(\frac{R^{(n)}_{+,i}(\theta)}{n+1}\Big)\Big)\Big(\Big(F^{-1}_{1+}\Big(\frac{R^{(n)}_{+,i}(\theta)}{n+1}\Big)\Big)^2 - \kappa(f_1)\Big).$$

Define the *cross-information coefficients*

$$\mathscr{I}(f_1,g_1) := \int_0^1 \phi_{f_1}\left(F_1^{-1}(u)\right)\phi_{g_1}\left(G_1^{-1}(u)\right)du,$$

$$\mathscr{J}(f_1, g_1) := \int_0^1 (F_1^{-1}(u))^2 \phi_{f_1}\left(F_1^{-1}(u)\right) \phi_{g_1}\left(G_1^{-1}(u)\right) du, \text{ and}$$

$$\mathscr{K}(f_1, g_1) := \int_0^1 (F_1^{-1}(u))^2 (G_1^{-1}(u))^2 \phi_{f_1}\left(F_1^{-1}(u)\right) \phi_{g_1}\left(G_1^{-1}(u)\right) du,$$

and denote by

$$\mathscr{F}_{f_1} := \{g_1 \in \mathscr{F}_0 | \mathscr{I}(f_1, g_1) < \infty,\ \mathscr{J}(f_1, g_1) < \infty, \text{ and } \mathscr{K}(f_1, g_1) < \infty\}$$

the class of densities for which those integrals exist and are finite.

Proposition 7.3 *Let $f_1 \in \mathscr{F}_1$. Then,*

(i) $\underset{\sim}{T}_{f_1}^{(n)}(\theta)$ is asymptotically normal, with mean zero under $\bigcup_{g_1 \in \mathscr{F}_0} \bigcup_{\sigma \in \mathbb{R}_0^+} \{\mathrm{P}_{\theta,\sigma,0;g_1}^{(n)}\}$, mean

$$\frac{\tau}{\gamma^{1/2}(f_1)} \left[\mathscr{K}(f_1, g_1) - \mathscr{J}(f_1, g_1)\kappa(g_1) - \mathscr{J}(g_1, f_1)\kappa(f_1) + \mathscr{I}(f_1, g_1)\kappa(f_1)\kappa(g_1)\right]$$

under $\bigcup_{\sigma \in \mathbb{R}_0^+} \{\mathrm{P}_{\theta,\sigma,n^{-1/2}\tau;g_1}^{(n)}\}$, $g_1 \in \mathscr{F}_{f_1}$, and variance one under both;

(ii) the sequence of tests rejecting the hypothesis $\mathscr{H}_\theta^{(n)} := \bigcup_{g_1 \in \mathscr{F}_0} \bigcup_{\sigma \in \mathbb{R}_0^+} \{\mathrm{P}_{\theta,\sigma,0;g_1}^{(n)}\}$ of symmetry with respect to θ whenever $\underset{\sim}{T}_{f_1}^{(n)}(\theta)$ exceeds the $(1-\alpha)$ standard normal quantile z_α is locally asymptotically most powerful, at asymptotic level α, against $\bigcup_{\xi>0} \bigcup_{\sigma \in \mathbb{R}_0^+} \{\mathrm{P}_{\theta,\sigma,\xi;f_1}^{(n)}\}$.

Only asymptotic critical values are reported in Part (ii) of the proposition, but exact (or simulated) ones of course also can be considered, as the test is entirely distribution-free. The two-sided version also readily follows.

7.3.2 Optimal Signed-Rank Tests of Symmetry: Unspecified Location

The unspecified-location case is more difficult and, to the best of our knowledge, so far only has been considered in Cassart et al. (2008). When θ is unspecified, a consistent estimator $\hat{\theta}$ has to be substituted for θ, yielding *aligned* signs $s_i(\hat{\theta})$ and *aligned* ranks $R_{+,i}^{(n)}(\hat{\theta})$. This substitution, however, has an impact on the distribution-freeness of $\underset{\sim}{T}_{f_1}^{(n)}$, and on its asymptotic distribution. That impact, as we shall see, requires a delicate asymptotic analysis.

As usual in this context, denoting by $\mathrm{P}_{\lambda}^{(n)}$ a sequence of probability measures indexed by some parameter λ, consider a sequence of estimators $\hat{\lambda}^{(n)}$ of λ satisfying, under $\mathrm{P}_{\lambda}^{(n)}$, the following assumptions:

(C1) (root-n consistency) $\hat{\lambda}^{(n)} - \lambda = O_{\mathrm{P}}(n^{-1/2})$, and
(C2) (local discreteness) the number of possible values of $\hat{\lambda}^{(n)}$ in balls with $O(n^{-1/2})$ radius centered at λ is bounded as $n \to \infty$.

An estimator $\lambda^{(n)}$ satisfying (C1) but not (C2) is easily discretized by letting, for some arbitrary $c > 0$, $\lambda_{\#}^{(n)} := (cn^{1/2})^{-1}\mathrm{sign}(\lambda^{(n)})\lceil cn^{1/2}|\lambda^{(n)}|\rceil$, which satisfies both (C1) and (C2). Subscripts $_{\#}$ in the sequel are used for estimators $(\hat{\theta}_{\#}, \hat{\sigma}_{\#}, \dots)$ satisfying (C1) and (C2). It should be noted, however, that (C2) has no implications in practice, where n is fixed, as the discretization constant c can be chosen arbitrarily large.

For all $\kappa \in \mathbb{R}_0^+$, define

$$\underset{\sim}{\Delta}_{f_1;3}^{(n)}(\kappa;\theta) := n^{-1/2}\sum_{i=1}^{n} s_i(\theta)\phi_{f_1}\Big(F_{1+}^{-1}\Big(\frac{R_{+,i}^{(n)}(\theta)}{n+1}\Big)\Big)\Big(\Big(F_{1+}^{-1}\Big(\frac{R_{+,i}^{(n)}(\theta)}{n+1}\Big)\Big)^2 - \kappa\Big);$$

for $\kappa = \kappa(f_1)$, $\underset{\sim}{\Delta}_{f_1;3}^{(n)}(\kappa;\theta)$ and $\underset{\sim}{\Delta}_{f_1;3}^{(n)}(\theta)$ coincide. The joint distribution, under $\mathrm{P}_{\theta,\sigma,0;g_1}^{(n)}$, of

$$\begin{pmatrix} \underset{\sim}{\Delta}_{f_1;3}^{(n)}(\kappa;\theta) \\ \Delta_{g_1;1}^{(n)}(\theta,\sigma,0) \end{pmatrix} = n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} \phi_{f_1}\Big(F_1^{-1}(G_1(Z_i^{(n)}(\theta,\sigma)))\Big)\Big(\Big(F_1^{-1}(G_1(Z_i^{(n)}(\theta,\sigma)))\Big)^2 - \kappa\Big) \\ \frac{1}{\sigma}\phi_{g_1}(Z_i(\theta,\sigma)) \end{pmatrix} + o_{\mathrm{P}}(1),$$

(an asymptotic representation similar to that of Part (i) of Proposition 7.2 clearly holds) is asymptotically normal, with mean zero and covariance matrix

$$\begin{pmatrix} \mathcal{Y}^{\kappa}(f_1) & \delta^{\kappa}(f_1,g_1) \\ \delta^{\kappa}(f_1,g_1) & \sigma^{-2}\mathscr{I}(g_1) \end{pmatrix}, \tag{7.10}$$

where

$$\mathcal{Y}^{\kappa}(f_1) := \mathscr{K}(f_1) - 2\kappa\mathscr{J}(f_1) + \kappa^2\mathscr{I}(f_1)$$

and

$$\delta^{\kappa}(f_1,g_1) := \sigma^{-1}(\mathscr{J}(f_1,g_1) - \kappa\mathscr{I}(f_1,g_1)).$$

It immediately follows from (7.10) and Le Cam's third Lemma that root-n perturbations of θ, hence also the replacement of θ by a root-n consistent estimator, do affect the asymptotic centering of $\underset{\sim}{\Delta}^{(n)}_{f_1;3}(\kappa;\theta)$—unless the covariance $\delta^{\kappa}(f_1,g_1)$ is zero, that is, unless

$$\kappa = \kappa(f_1,g_1) := \mathscr{J}(f_1,g_1)/\mathscr{I}(f_1,g_1).$$

Let $\underset{\sim}{\kappa}^{(n)}(f_1;\theta)$ be a consistent (under $\mathrm{P}^{(n)}_{\theta,\sigma,0;g_1}$) estimator of $\kappa(f_1,g_1)$. Since the mapping from κ to the values of $\underset{\sim}{\Delta}^{(n)}_{f_1;3}(\kappa;\theta)$ is continuous, $\underset{\sim}{\Delta}^{(n)}_{f_1;3}(\underset{\sim}{\kappa}^{(n)}(f_1;\theta);\theta)$ also can be expected to be asymptotically insensitive to root-n perturbations of θ. This is indeed the case, and the same reasoning as in Sect. 3.2.2 of Cassart et al. (2011) yields, under $\mathrm{P}^{(n)}_{\theta,\sigma,0;g_1}$, the asymptotic equivalence

$$\underset{\sim}{\Delta}^{(n)}_{f_1;3}(\underset{\sim}{\kappa}^{(n)}(f_1;\hat{\theta}_{\#});\hat{\theta}_{\#}) - \underset{\sim}{\Delta}^{(n)}_{f_1;3}(\underset{\sim}{\kappa}^{(n)}(f_1;\theta);\theta) = o_{\mathrm{P}}(1)$$

for any estimator $\hat{\theta}_{\#}$ of θ satisfying (C1) and (C2) and any density g_1 in

$$\mathscr{F}^{*}_{f_1} := \{g_1 \in \mathscr{F}_0 \mid \mathscr{I}(g_1) < \infty,\ \mathscr{I}(f_1,g_1) < \infty,\ \text{and } \mathscr{J}(f_1,g_1) < \infty\}.$$

Obtaining a consistent estimator $\underset{\sim}{\kappa}^{(n)}(f_1;\theta)$ of $\underset{\sim}{\kappa}(f_1,g_1)$, which is a ratio of expected values, taken under unspecified density g, of variables that themselves depend on g, however, is delicate. A general method is proposed in Cassart et al. (2010), which we describe in Sect. 7.4. Defining

$$\underset{\sim}{T}^{(n)*}_{f_1}(\theta) := \Big(n\underset{\sim}{\gamma}^{(n)*}_{f_1}\Big)^{-1/2} \times \sum_{i=1}^{n} s_i(\theta)\phi_{f_1}\Big(F^{-1}_{1+}\Big(\frac{R^{(n)}_{+,i}(\theta)}{n+1}\Big)\Big)\Big(\Big(F^{-1}_{1+}\Big(\frac{R^{(n)}_{+,i}(\theta)}{n+1}\Big)\Big)^2 - \underset{\sim}{\kappa}^{(n)}(f_1;\theta)\Big), \tag{7.11}$$

where (convergence under $\mathrm{P}^{(n)}_{\theta,\sigma,0;g_1}$, as $n\to\infty$)

$$\begin{aligned}\underset{\sim}{\gamma}^{(n)*}_{f_1} &:= n^{-1}\sum_{r=1}^{n}\phi^2_{f_1}\Big(F^{-1}_{1+}\Big(\frac{r}{n+1}\Big)\Big)\Big(\Big(F^{-1}_{1+}\Big(\frac{r}{n+1}\Big)\Big)^2 - \underset{\sim}{\kappa}^{(n)}(f_1;\theta)\Big)^2 \\ &= \mathscr{K}(f_1) - 2\underset{\sim}{\kappa}(f_1,g_1)\mathscr{J}(f_1) + \underset{\sim}{\kappa}^2(f_1,g_1)\mathscr{I}(f_1) + o(1),\end{aligned}$$

we thus have established the following result.

Proposition 7.4 *Let $f_1 \in \mathscr{F}_1$. Then,*

(i) $\underset{\sim}{T}^{(n)*}_{f_1}(\hat{\theta}_\#)$ is asymptotically normal, with mean zero under the null hypothesis $\bigcup_{g_1 \in \underset{\sim}{\mathscr{F}}^*_{f_1}} \bigcup_{\theta \in \mathbb{R}} \bigcup_{\sigma \in \mathbb{R}_0^+} \{\mathrm{P}^{(n)}_{\theta,\sigma,0;g_1}\}$ of symmetry with respect to unspecified θ, mean

$$\tau \frac{\mathscr{K}(f_1, g_1) - \mathscr{J}(g_1, f_1)\underset{\sim}{\mathscr{K}}(f_1, g_1)}{\left(\mathscr{K}(f_1) - 2\underset{\sim}{\mathscr{K}}(f_1, g_1)\mathscr{J}(f_1) + \underset{\sim}{\mathscr{K}}^2(f_1, g_1)\mathscr{I}(f_1)\right)^{1/2}}$$

under the local alternative $\mathrm{P}^{(n)}_{\theta,\sigma,n^{-1/2}\tau;g_1}$, where $g_1 \in \underset{\sim}{\mathscr{F}}^*_{f_1}$, and variance one under both;

(ii) the sequence of tests rejecting the null hypothesis $\bigcup_{g_1 \in \underset{\sim}{\mathscr{F}}^*_{f_1}} \bigcup_{\theta \in \mathbb{R}} \bigcup_{\sigma \in \mathbb{R}_0^+} \{\mathrm{P}^{(n)}_{\theta,\sigma,0;g_1}\}$ of symmetry with respect to unspecified θ whenever $\underset{\sim}{T}^{(n)*}_{f_1}(\hat{\theta}_\#)$ exceeds the $(1-\alpha)$ standard normal quantile z_α is locally asymptotically most powerful, at asymptotic level α, against $\bigcup_{\xi>0} \bigcup_{\theta \in \mathbb{R}} \bigcup_{\sigma \in \mathbb{R}_0^+} \{\mathrm{P}^{(n)}_{\theta,\sigma,\xi;f_1}\}$.

7.3.3 The van der Waerden, Wilcoxon, and Laplace Tests of Symmetry

Important particular cases of (7.9) and (7.11) are the Laplace (sign test or double-exponential scores), Wilcoxon (logistic scores), and van der Waerden (normal scores) tests, which are optimal at double-exponential, logistic, and normal distributions, respectively.

The van der Waerden tests are based on $f_1 = \phi_1$, with

$$F^{-1}_{1+}(u) = \phi_{f_1}(F^{-1}_{1+}(u)) = a^{-1/2}\Phi^{-1}\Big(\frac{u+1}{2}\Big),$$

where Φ is the standard normal distribution function. The specified-location test statistic (7.9) then reduces to

$$\underset{\sim}{T}^{(n)}_{\mathrm{vdW}}(\theta) := \Big(n\underset{\sim}{\mathcal{Y}}^{(n)}_{\phi_1}\Big)^{-1/2} \times \sum_{i=1}^{n} s_i(\theta)\Phi^{-1}\Big(\frac{n+1+R^{(n)}_{+,i}(\theta)}{2(n+1)}\Big)\Bigg(\Big(\Phi^{-1}\Big(\frac{n+1+R^{(n)}_{+,i}(\theta)}{2(n+1)}\Big)\Big)^2 - 3\Bigg),$$

where

$$\underset{\sim}{\mathcal{Y}}^{(n)}_{\phi_1} := n^{-1}\sum_{r=1}^{n}\left(\Phi^{-1}\Big(\frac{n+1+r}{2(n+1)}\Big)\right)^2\left(\Big(\Phi^{-1}\Big(\frac{n+1+r}{2(n+1)}\Big)\Big)^2 - 3\right)^2.$$

The corresponding unspecified-location test statistic (7.11) takes the form

$$\underset{\sim}{T}^{(n)*}_{\mathrm{vdW}}(\hat{\theta}) := \left(n\underset{\sim}{\gamma}^{(n)*}_{\phi_1}\right)^{-1/2} \sum_{i=1}^{n} s_i(\hat{\theta}) \Phi^{-1}\left(\frac{n+1+R^{(n)}_{+,i}(\hat{\theta})}{2(n+1)}\right) \times \left(\left(\Phi^{-1}\left(\frac{n+1+R^{(n)}_{+,i}(\hat{\theta})}{2(n+1)}\right)\right)^2 - \underset{\sim}{\kappa}^{(n)}(\phi_1;\hat{\theta})\right),$$

where

$$\underset{\sim}{\gamma}^{(n)*}_{\phi_1} := n^{-1} \sum_{r=1}^{n} \left(\Phi^{-1}\left(\frac{n+1+r}{2(n+1)}\right)\right)^2 \left(\left(\Phi^{-1}\left(\frac{n+1+r}{2(n+1)}\right)\right)^2 - \underset{\sim}{\kappa}^{(n)}(\phi_1;\hat{\theta})\right)^2.$$

In the Wilcoxon case (logistic density), one easily checks that

$$F^{-1}_{1+}(u) = b^{-1/2} \log \frac{1+u}{1-u} \quad \text{and} \quad \phi_{f_1}(F^{-1}_{1+}(u)) = b^{-1/2}u.$$

Therefore, (7.9) and (7.11) reduce to

$$\underset{\sim}{T}^{(n)}_{\mathrm{W}}(\theta) := \left(n\underset{\sim}{\gamma}^{(n)}_{f_{\mathrm{Log}}}\right)^{-1/2} \sum_{i=1}^{n} s_i(\theta) R^{(n)}_{+,i}(\theta) \left(\left(\log \frac{n+1+R^{(n)}_{+,i}(\theta)}{n+1-R^{(n)}_{+,i}(\theta)}\right)^2 - \frac{12+\pi^2}{3}\right),$$

and

$$\underset{\sim}{T}^{(n)*}_{\mathrm{W}}(\hat{\theta}) := \left(n\underset{\sim}{\gamma}^{(n)*}_{f_{\mathrm{Log}}}\right)^{-1/2} \sum_{i=1}^{n} s_i(\hat{\theta}) R^{(n)}_{+,i}(\hat{\theta}) \left(\left(\log \frac{n+1+R^{(n)}_{+,i}(\hat{\theta})}{n+1-R^{(n)}_{+,i}(\hat{\theta})}\right)^2 - \underset{\sim}{\kappa}^{(n)}(f_{\mathrm{Log}};\hat{\theta})\right),$$

where

$$\underset{\sim}{\gamma}^{(n)}_{f_{\mathrm{Log}}} := n^{-1} \sum_{r=1}^{n} r^2 \left(\left(\log \frac{n+1+r}{n+1-r}\right)^2 - \frac{12+\pi^2}{3}\right)^2$$

and

$$\underset{\sim}{\gamma}^{(n)*}_{f_{\mathrm{Log}}} := n^{-1} \sum_{r=1}^{n} r^2 \left(\left(\log \frac{n+1+r}{n+1-r}\right)^2 - \underset{\sim}{\kappa}^{(n)}(f_{\mathrm{Log}};\hat{\theta})\right)^2,$$

respectively.

As for the Laplace-score version of (7.9), it is associated with the double-exponential density $f_1 = f_{\mathscr{L}}$. One easily obtains

$$F_{1+}^{-1}(u) = -d\log(1-u) \quad \text{and} \quad \phi_{f_1}(F_{1+}^{-1}(u)) = 1/d,$$

hence

$$\underset{\sim}{T}_{\mathscr{L}}^{(n)}(\theta) := \left(n\underset{\sim}{\gamma}_{f_{\mathscr{L}}}^{(n)}\right)^{-1/2} \sum_{i=1}^{n} s_i(\theta)\left(\left(\log\left(1 - \frac{R_{+,i}^{(n)}(\theta)}{n+1}\right)\right)^2 - 2\right),$$

where

$$\underset{\sim}{\gamma}_{f_{\mathscr{L}}}^{(n)} := n^{-1}\sum_{r=1}^{n}\left(\left(\log\left(1 - \frac{r}{n+1}\right)\right)^2 - 2\right)^2.$$

The unspecified-location test statistic (7.11) is derived along the same lines as previously, yielding

$$\underset{\sim}{T}_{\mathscr{L}}^{(n)*}(\hat{\theta}) := \left(n\underset{\sim}{\gamma}_{f_{\mathscr{L}}}^{(n)*}\right)^{-1/2} \sum_{i=1}^{n} s_i(\hat{\theta})\left(\left(\log\left(1 - \frac{R_{+,i}^{(n)}(\hat{\theta})}{n+1}\right)\right)^2 - \underset{\sim}{\kappa}^{(n)}(f_{\mathscr{L}};\hat{\theta})\right),$$

with

$$\underset{\sim}{\gamma}_{f_{\mathscr{L}}}^{(n)*} := n^{-1}\sum_{r=1}^{n}\left(\left(\log\left(1 - \frac{r}{n+1}\right)\right)^2 - \underset{\sim}{\kappa}^{(n)}(f_{\mathscr{L}};\hat{\theta})\right)^2.$$

7.3.4 *Asymptotic Relative Efficiencies*

The asymptotic shifts provided in Propositions 7.3 and 7.4, together with their pseudo-Gaussian counterparts in Propositions 3.2 and 3.5 of Cassart et al. (2011) allow us to compute ARE values for the tests based on $\underset{\sim}{T}_{f_1}^{(n)}(\theta)$ and $\underset{\sim}{T}_{f_1}^{(n)*}(\hat{\theta}_{\#})$ with respect to their classical counterparts, based on $m_3^{(n)}(\theta)$ [or $S_1^{(n)}$; see (7.1)] and $b_1^{(n)}$ [or $S_2^{(n)}$; see (7.2)], respectively. Those ARE values are obtained as the squared ratios of the corresponding local shifts, for various densities g_1. The pseudo-Gaussian tests, hence also our ARE values, require finite sixth-order moments. But signed-rank tests of course remain valid without such assumption and, whenever g_1 has infinite moment of order six, the asymptotic relative efficiency under g_1 of any signed-rank test with respect to its pseudo-Gaussian competitor can be considered as infinite.

Proposition 7.5 *Let $f_1 \in \mathscr{F}_1$; denoting by μ_k the moment of order k of g_1, assume that $\mu_6 < \infty$.*

(i) The asymptotic relative efficiency under $g_1 \in \mathscr{F}_{\phi_1} \cap \mathscr{Z}_{f_1}$ of the specified-location signed-rank test based on $\underset{\sim}{T}_{f_1}^{(n)}(\theta)$ with respect to the classical procedure based on $m_3^{(n)}(\theta)$ [see (7.1)] is

$$ARE_{g_1}\left(\underset{\sim}{T}_{f_1}^{(n)}(\theta)/m_3^{(n)}(\theta)\right) = \frac{\left(\mathscr{K}(f_1,g_1) - \mathscr{J}(f_1,g_1)\kappa(g_1) - \mathscr{J}(g_1,f_1)\kappa(f_1) + \mathscr{I}(f_1,g_1)\kappa(f_1)\kappa(g_1)\right)^2}{\gamma(f_1)\left(5\mu_4 - 3\kappa(g_1)\mu_2\right)^2/\mu_6}.$$

(ii) The asymptotic relative efficiency under $g_1 \in \mathscr{F}_{\phi_1} \cap \mathscr{Z}_{f_1}$ of the specified-location signed-rank test based on $\underset{\sim}{T}_{f_1}^{(n)}(\theta)$ with respect to the classical procedure based on $b_1^{(n)}$ [see (7.2)] is

$$ARE_{g_1}\left(\underset{\sim}{T}_{f_1}^{(n)}(\theta)/b_1^{(n)}\right) = \frac{\left(\mathscr{K}(f_1,g_1) - \mathscr{J}(g_1,f_1)\kappa(g_1) - \mathscr{J}(g_1,f_1)\kappa(f_1) + \mathscr{I}(f_1,g_1)\kappa(f_1)\kappa(g_1)\right)^2}{\gamma(f_1)\left(5\mu_4 - 9\mu_2^2\right)^2/\left(\mu_6 - 6\mu_2\mu_4 + 9\mu_2^3\right)}.$$

(iii) The asymptotic relative efficiency under $g_1 \in \mathscr{F}_{\phi_1} \cap \mathscr{Z}_{f_1}^*$ of the unspecified-location signed-rank test based on $\underset{\sim}{T}_{f_1}^{(n)*}(\hat{\theta}_\#)$ with respect to the classical procedure based on $b_1^{(n)}$ [see (7.2)] is

$$ARE_{g_1}\left(\underset{\sim}{T}_{f_1}^{(n)*}(\hat{\theta}_\#)/b_1^{(n)}\right) = \frac{\left(\mathscr{K}(f_1,g_1) - \mathscr{J}(g_1,f_1)\underset{\sim}{\kappa}(f_1,g_1)\right)^2\left(\mu_6 - 6\mu_2\mu_4 + 9\mu_2^3\right)}{\left(\mathscr{K}(f_1) - 2\underset{\sim}{\kappa}(f_1,g_1)\mathscr{J}(f_1) + \underset{\sim}{\kappa}^2(f_1,g_1)\mathscr{I}(f_1)\right)\left(5\mu_4 - 9\mu_2^2\right)^2}.$$

Numerical values of those AREs, under Student (6.5, 8, 10, and 20 degrees of freedom), normal, power-exponential (exponents 2, 3, and 5), and logistic densities are displayed in Tables 7.1 and 7.2. Those values are quite high, particularly so under heavy tails (see, for instance, the Student density with 6.5 degrees of freedom). Also, the AREs of the van der Waerden tests are uniformly larger than or equal to one. The tests with power-exponential scores, however, are not performing well under Student and logistic densities. On the other hand, under the power-exponential density with exponent 2, the classical procedure based on $m_3^{(n)}(\theta)$ has no power at all, yielding infinite ARE values (the shift in the denominator is zero) in Table 7.1.

Table 7.1 AREs, under Student ($f_{t_{6.5}}, f_{t_8}, f_{t_{10}}$ and $f_{t_{20}}$), normal (ϕ_1), power-exponential ($f_{\mathscr{E}_2}, f_{\mathscr{E}_3}$, $f_{\mathscr{E}_5}$, with exponents 2, 3, and 5), and logistic (f_{Log}) densities, of various signed-rank tests (based on Student, van der Waerden, power-exponential, and Wilcoxon scores), with respect to the pseudo-Gaussian test (based on $b_1^{(n)}$, see Proposition 7.5(ii); first line) and with respect to the classical test of skewness (based on $m_3^{(n)}$, see Proposition 7.5(i); second line), for testing symmetry about a specified location θ

	Actual density g_1								
Score f_1	$f_{t_{6.5}}$	f_{t_8}	$f_{t_{10}}$	$f_{t_{20}}$	ϕ_1	$f_{\mathscr{E}_2}$	$f_{\mathscr{E}_3}$	$f_{\mathscr{E}_5}$	f_{Log}
$f_{t_{6.5}}$	4.6923	1.8304	1.3556	1.0350	0.9374	0.9593	1.0669	0.9055	1.1843
	3.9000	1.8338	1.6019	1.7313	2.3436	∞	3.6063	0.3666	1.3948
f_{t_8}	4.6848	1.8333	1.3618	1.0462	0.9542	1.0170	1.1670	1.0646	1.1849
	3.8938	1.8367	1.6091	1.7501	2.3855	∞	3.9446	0.4309	1.3955
$f_{t_{10}}$	4.6648	1.8308	1.3636	1.0540	0.9679	1.0733	1.2707	1.2459	1.1826
	3.8771	1.8342	1.6113	1.7630	2.4197	∞	4.2951	0.5044	1.3928
$f_{t_{20}}$	4.5765	1.8074	1.3543	1.0612	0.9904	1.2008	1.5258	1.7512	1.1670
	3.8037	1.8108	1.6003	1.7752	2.4759	∞	5.1574	0.7089	1.3745
ϕ_1	4.3988	1.7494	1.3199	1.0510	1.0000	1.3402	1.8394	2.4753	1.1304
	3.6560	1.7526	1.5596	1.7580	2.5000	∞	6.2175	1.0020	1.3313
$f_{\mathscr{E}_2}$	2.5240	1.0455	0.8207	0.7145	0.7515	1.7834	3.3911	7.2764	0.6740
	2.0978	1.0475	0.9698	1.1952	1.8788	∞	11.4624	2.9454	0.7938
$f_{\mathscr{E}_3}$	1.3575	0.5802	0.4699	0.4391	0.4988	1.6399	3.6878	9.7508	0.3776
	1.1283	0.5812	0.5552	0.7345	1.2470	∞	12.4654	3.9471	0.4448
$f_{\mathscr{E}_5}$	0.3853	0.1770	0.1541	0.1685	0.2245	1.1766	3.2606	11.0283	0.1208
	0.3202	0.1773	0.1820	0.2819	0.5611	∞	11.0214	4.4642	0.1423
f_{Log}	4.6839	1.8311	1.3593	1.0439	0.9528	1.0132	1.1739	1.1232	1.1864
	3.8930	1.8345	1.6062	1.7462	2.3820	∞	3.9680	0.4547	1.3973

Table 7.2 AREs, under Student ($f_{t_{6.5}}, f_{t_8}, f_{t_{10}}$ and $f_{t_{20}}$), normal (ϕ_1), power-exponential ($f_{\mathscr{E}_2}, f_{\mathscr{E}_3}$, $f_{\mathscr{E}_5}$, with exponents 2, 3, and 5), and logistic (f_{Log}) densities, of various signed-rank tests (based on Student, van der Waerden, power-exponential, and Wilcoxon scores), with respect to the pseudo-Gaussian test [based on $b_1^{(n)}$, see Proposition 7.5(iii)], for testing symmetry with respect to unspecified location θ

	Actual density g_1								
Score f_1	$f_{t_{6.5}}$	f_{t_8}	$f_{t_{10}}$	$f_{t_{20}}$	ϕ_1	$f_{\mathscr{E}_2}$	$f_{\mathscr{E}_3}$	$f_{\mathscr{E}_5}$	f_{Log}
$f_{t_{6.5}}$	4.6923	2.7856	1.3628	1.0580	0.9902	1.3769	2.0874	3.7623	1.1853
f_{t_8}	4.6915	1.8333	1.3634	1.0591	0.9919	1.3974	2.1401	3.9103	1.1850
$f_{t_{10}}$	4.6895	2.7846	1.3636	1.0601	0.9938	1.4173	2.1906	4.0506	1.1848
$f_{t_{20}}$	4.6803	2.7795	1.3623	1.0612	0.9978	1.4627	2.3052	4.3671	1.1839
ϕ_1	4.6547	2.7642	1.3558	1.0589	1	1.5129	2.4359	4.7336	1.1799
$f_{\mathscr{E}_2}$	4.1723	1.6254	1.2084	0.9448	0.8993	1.7834	3.5036	8.4598	1.0536
$f_{\mathscr{E}_3}$	3.7109	1.4365	1.0629	0.8255	0.7833	1.7052	3.6878	10.1646	0.9311
$f_{\mathscr{E}_5}$	3.0659	1.1736	0.8610	0.6594	0.6200	1.4453	3.4284	11.0283	0.7591
f_{Log}	4.6878	2.7828	1.3618	1.0585	0.9926	1.3728	2.0678	3.6975	1.1864

The van der Waerden tests thus appear as a very attractive alternative to the classical tests—all the more so that their validity does not require the actual density g to have finite moments of order 6; recall, however, that $g_1 \in \mathscr{F}^*_{f_1}$ (here, $\mathscr{F}^*_{\phi_1}$) is still needed in the unspecified-location case.

7.4 Estimation of Cross-Information Quantities

7.4.1 *Consistent Estimation of* $\mathscr{I}(f_1, g_1)$ *and* $\mathscr{J}(f_1, g_1)$

Implementing the unspecified-location rank-based tests of Sect. 7.3.2 thus requires consistent estimation of $\kappa(f_1, g_1)$, that is, consistent estimation of the cross-information quantities $\mathscr{I}(f_1, g_1)$ and $\mathscr{J}(f_1, g_1)$. The cross-information for location $\mathscr{I}(f_1, g_1)$ is a familiar quantity in classical rank-based inference. It explicitly appears, indeed, in the asymptotic powers of traditional rank and signed-rank tests for location, and in the asymptotic variance of the corresponding R-estimators. In a different context (R-estimation of shape), its counterpart also plays a central role in the construction of one-step R-estimators (Hallin et al. 2006).

Estimating $\mathscr{I}(f_1, g_1)$, however, is all but straightforward. No empirical version of this expectation is available, as it involves the unknown score ϕ_{g_1} associated with the unspecified density g_1. Various methods have been proposed for estimating $\mathscr{I}(f_1, g_1)$ in connection with R-estimation. Some of them (Lehmann 1963; Sen 1966) involve comparisons of lengths of confidence intervals. Some others (Kraft and van Eeden (1972); Antille (1974), or Jurečková and Sen (1996, page 321)) rely on the asymptotic linearity property of rank statistics. More elaborated approaches involve kernel estimates of g_1—hence cannot be expected to perform well under small and moderate sample sizes. Such kernel methods have been considered, for Wilcoxon scores, by Schweder (1975) (see also Cheng and Serfling 1981; Bickel and Ritov 1988; Fan 1991) and, in a more general setting, in Sect. 4.5 of Koul (2002).

All these methods, however, involve quite arbitrary choices (choice of a confidence level for confidence intervals; choice of an arbitrary $O(n^{-1/2})$ perturbation for the method based on asymptotic linearity; choice of a kernel and a bandwidth for the estimation of g_1). Although they do not affect consistency and asymptotic efficiency, such choices may have a dramatic impact on finite sample results. As for kernel methods, they require large sample sizes and are kind of antinomic to the spirit of rank-based methods: if densities are to be estimated, indeed, using them all the way by inserting estimated scores into the parametric tests of Cassart et al. (2011) seems more coherent than considering ranks.

The approach proposed in Hallin et al. (2006) is of an entirely different nature; the basic idea consists in solving a local linearized likelihood equation. In the present setting, this estimator of $\mathscr{I}(f_1, g_1)$ is constructed as follows. Denoting by $\hat{\theta}$

and $\widehat{\sigma}$ root-n consistent (under $\mathrm{P}^{(n)}_{\theta,\sigma,0;g_1}$, $g_1 \in \mathscr{F}^*_{f_1}$) estimators of θ and σ, respectively (the median and the median of absolute deviations constitute appropriate choices), by $\hat{\theta}_\#$ and $\widehat{\sigma}_\#$ their discretized versions, and by $\underset{\sim}{\Delta}^{(n)}_{f_1;1;\#}$ a discretized version of $\underset{\sim}{\Delta}^{(n)}_{f_1;1}$, let, for any $\beta > 0$,

$$\underset{\sim}{\theta}^{(n)}_*(\beta) := \hat{\theta}_\# + n^{-1/2}\beta\widehat{\sigma}^2_\#\underset{\sim}{\Delta}^{(n)}_{f_1;1;\#}(\hat{\theta}_\#). \tag{7.12}$$

Choosing a further arbitrary discretization constant $c > 0$, put $\beta_\ell := \ell/c$, $\ell \in \mathbb{N}$, and define

$$\beta_1^- := \min\{\beta_\ell \,|\, \underset{\sim}{\Delta}^{(n)}_{f_1;1;\#}(\underset{\sim}{\theta}^{(n)}_*(\beta_{\ell+1}))\,\underset{\sim}{\Delta}^{(n)}_{f_1;1;\#}(\hat{\theta}_\#) < 0\}, \qquad \beta_1^+ := \beta_1^- + \frac{1}{c},$$

and

$$\beta_1^* := \beta_1^- + \frac{1}{c}\frac{\underset{\sim}{\Delta}^{(n)}_{f_1;1;\#}(\underset{\sim}{\theta}^{(n)}_*(\beta_1^-))}{\underset{\sim}{\Delta}^{(n)}_{f_1;1;\#}(\underset{\sim}{\theta}^{(n)}_*(\beta_1^-)) - \underset{\sim}{\Delta}^{(n)}_{f_1;1;\#}(\underset{\sim}{\theta}^{(n)}_*(\beta_1^+))}.$$

Then,

$$\mathscr{I}^{(n)}(f_1) := (\beta_1^*)^{-1} = \mathscr{I}(f_1, g_1) + o_{\mathrm{P}}(1)$$

under $\mathrm{P}^{(n)}_{\theta,\sigma,0;g_1}$, $g_1 \in \mathscr{F}^*_{f_1}$, as $n \to \infty$. Moreover, $\underset{\sim}{\theta}^{(n)}_{f_1} := \underset{\sim}{\theta}^{(n)}_*(\beta_1^*)$ is an efficient (at $\mathrm{P}^{(n)}_{\theta,\sigma,0;f_1}$—efficiency here is in the parametric sense) R-estimator of θ. A proof for this can be obtained by parallelling that of Sect. 4.2 in Hallin et al. (2006). The same claim, however, also follows from a more general result by Cassart et al. (2010) which we also need for the estimation of $\mathscr{J}(f_1, g_1)$.

The estimation method just described for $\mathscr{I}(f_1, g_1)$ indeed does not apply to $\mathscr{J}(f_1, g_1)$ which, contrary to $\mathscr{I}(f_1, g_1)$, is not associated with any optimal one-step R-estimation procedure (it does not follow as a coefficient of the covariance, under $\mathrm{P}^{(n)}_{\theta,\sigma,0;g_1}$, of any component $\underset{\sim}{\Delta}^{(n)}_{f_1;\ell}(\boldsymbol{\vartheta})$ of the rank-based version of the central sequence with the corresponding component $\Delta^{(n)}_{g_1;\ell}(\boldsymbol{\vartheta})$ of $\boldsymbol{\Delta}^{(n)}_{g_1}(\boldsymbol{\vartheta})$). Fortunately, Proposition 2.1 of Cassart et al. (2010) applies, yielding the desired estimator.

That proposition requires the existence of a rank-based statistic $\underset{\sim}{S}^{(n)}(\theta)$ satisfying, under $\mathrm{P}^{(n)}_{\theta,\sigma,0;g_1}$ and for all σ, the following two conditions (Assumption (A) of Cassart et al. 2010).

(D1) The sequence $\underset{\sim}{S}^{(n)}(\theta)$, $n \in \mathbb{N}$ is tight and bounded away from zero, that is, for all $\varepsilon > 0$, there exist δ_ε, M_ε and N_ε such that, for all $n > N_\varepsilon$,

$$\mathrm{P}^{(n)}_{\theta,\sigma,0;g_1}[\delta_\varepsilon \le |\underset{\sim}{S}^{(n)}(\theta)| \le M_\varepsilon] \ge 1 - \varepsilon. \tag{7.13}$$

(D2) For any θ, there exists $\Upsilon(\theta) \neq 0$ such that, for any bounded sequence $t^{(n)}$,

$$\underset{\sim}{S}^{(n)}(\theta + n^{-1/2}t^{(n)}) = \underset{\sim}{S}^{(n)}(\theta) - t^{(n)}\mathscr{J}(g)\Upsilon^{-1}(\theta) + o_{\mathrm{P}}(1) \quad \text{as } n \to \infty, \tag{7.14}$$

where the mapping $\theta \mapsto \Upsilon(\theta)$ is continuous.

In the present context, we propose

$$\underset{\sim}{S}^{(n)}(\theta) := \underset{\sim}{S}^{(n)}_{f_1}(\theta) := n^{-1/2}\sum_{i=1}^{n} s_i^{(n)}(\theta)\phi_{f_1}\Big(F_{1+}^{-1}\Big(\frac{R_{+,i}^{(n)}(\theta)}{n+1}\Big)\Big)\Big(F_{1+}^{-1}\Big(\frac{R_{+,i}^{(n)}(\theta)}{n+1}\Big)\Big)^2.$$

It follows from the classical Hájek results for linear signed-rank statistics (see, e.g., Chapter 3 of Puri and Sen 1985) that $\underset{\sim}{S}^{(n)}_{f_1}(\theta)$ is asymptotically centered normal, so that (7.13) is satisfied, while (7.14) holds, in view of asymptotic linearity, for $g_1 \in \mathscr{F}^*_{f_1}$, with $\Upsilon(\theta) = \sigma\mathscr{J}^{-1}(f_1, g_1)$.

As in the estimation of $\mathscr{I}(f_1, g_1)$, root-n consistent preliminary estimators $\hat{\theta}$ and $\widehat{\sigma}$ of θ and σ are required as well; $\hat{\theta}$ moreover should be such that (Assumption (B) of Cassart et al. 2010), under $\mathrm{P}^{(n)}_{\theta,\sigma,0;g_1}$ and for all σ, $\underset{\sim}{S}^{(n)}_{f_1}(\hat{\theta})$ is not $o_{\mathrm{P}}(1)$—a condition which is trivially satisfied by the empirical median. All assumptions of Proposition 2.1 of Cassart et al. (2010) then hold, which entails the desired consistency of the estimator we now describe.

Proceeding as in (7.12) above, let (with the usual discretized versions $\hat{\theta}_{\#}$ and $\widehat{\sigma}_{\#}$)

$$\underset{\sim}{\theta}^{(n)}_{**}(\beta) := \hat{\theta}_{\#} + n^{-1/2}\beta\widehat{\sigma}_{\#}\underset{\sim}{S}^{(n)}_{f_1;\#}(\hat{\theta}_{\#}),$$

$$\beta_2^- := \min\{\beta_\ell \mid \underset{\sim}{S}^{(n)}_{f_1;\#}(\underset{\sim}{\theta}^{(n)}_{**}(\beta_{\ell+1}))\underset{\sim}{S}^{(n)}_{f_1;\#}(\hat{\theta}_{\#}) < 0\}, \quad \text{and} \quad \beta_2^+ := \beta_2^- + \frac{1}{c}.$$

Defining

$$\beta_2^* := \beta_2^- + \frac{1}{c}\frac{\underset{\sim}{S}^{(n)}_{f_1;\#}(\underset{\sim}{\theta}^{(n)}_{**}(\beta_2^-))}{\underset{\sim}{S}^{(n)}_{f_1;\#}(\underset{\sim}{\theta}^{(n)}_{**}(\beta_2^-)) - \underset{\sim}{S}^{(n)}_{f_1;\#}(\underset{\sim}{\theta}^{(n)}_{**}(\beta_2^+))},$$

the estimator $\mathscr{J}^{(n)}(f_1) := (\beta_2^*)^{-1}$ then is such that

$$\mathscr{J}^{(n)}(f_1) = \mathscr{J}(f_1, g_1) + o_{\mathrm{P}}(1) \quad \text{under } \mathrm{P}^{(n)}_{\theta,\sigma,0;g_1},\ g_1 \in \mathscr{F}^*_{f_1},\ \text{as } n \to \infty,$$

as desired.

7.4.2 Practical Implementation

As usual, all discretizations in the construction of $\mathscr{I}^{(n)}(f_1)$ and $\mathscr{J}^{(n)}(f_1)$ are required for the purpose of asymptotic statements, but can be dispensed with in applications, where n remains fixed. The practical versions of $\mathscr{I}^{(n)}(f_1)$ and $\mathscr{J}^{(n)}(f_1)$ therefore are

$$\mathscr{I}^{(n)}(f_1) := (\beta_1^*)^{-1} \quad \text{and} \quad \mathscr{J}^{(n)}(f_1) := (\beta_2^*)^{-1},$$

respectively, where

$$\beta_1^* := \inf\left\{\beta > 0 \,|\, \underset{\sim}{\Delta}_{f_1;1}^{(n)}(\hat{\theta} + n^{-1/2}\beta\widehat{\sigma}^2\underset{\sim}{\Delta}_{f_1;1}^{(n)}(\hat{\theta}))\underset{\sim}{\Delta}_{f_1;1}^{(n)}(\hat{\theta}) \leq 0\right\}$$

and

$$\beta_2^* := \inf\left\{\beta > 0 \,|\, \underset{\sim}{S}_{f_1}^{(n)}(\hat{\theta} + n^{-1/2}\beta\widehat{\sigma}\underset{\sim}{S}_{f_1}^{(n)}(\hat{\theta}))\underset{\sim}{S}_{f_1}^{(n)}(\hat{\theta}) \leq 0\right\},$$

and follow from adopting "large" values of the discretizing constants (or, letting c tend to infinity). The ratio $\mathscr{J}^{(n)}(f_1)/\mathscr{I}^{(n)}(f_1)$ then provides the estimator $\underset{\sim}{\kappa}^{(n)}(f_1;\theta)$ of $\underset{\sim}{\kappa}(f_1, g_1)$ required in the definition (7.11) of the test statistic of Proposition 7.4.

Acknowledgements Marc Hallin and Davy Paindaveine acknowledge the support of the IAP research network grant P7/06 of the Belgian federal government (Belgian Science Policy (BELSPO)). The research of Davy Paindaveine is also supported by an A.R.C. contract from the Communauté Française de Belgique, that of Marc Hallin by a Crédit aux Charcheurs of the Fonds de la Recherche Scientifique-FNRS.

References

Antille, A.: A linearized version of the Hodges-Lehmann estimator. Ann. Stat. **2**, 1308–1313 (1974)

Azzalini, A., Capitanio, A.: Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. J. R. Stat. Soc. Ser. B **65**, 367–389 (2003)

Bickel, P.J., Ritov, Y.: Estimating integrated squared density derivatives. Sankhya A **50**, 381–393 (1988)

Cassart, D., Hallin, M., Paindaveine, D.: Optimal detection of Fechner-asymmetry. J. Stat. Plan. Inference **138**, 2499–2525 (2008)

Cassart, D., Hallin, M., Paindaveine, D.: On the estimation of cross-information quantities in R-estimation. In: Antoch, J., Hušková, M., Sen, P.K. (eds.). Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in Honor of Professor Jana Jurečková. I.M.S. Monographs-Lecture Notes, pp. 35–45. Institute of Mathematical Statistics, Beachwood (2010)

Cassart, D., Hallin, M., Paindaveine, D.: A class of optimal tests for symmetry based on local Edgeworth approximations. Bernoulli **17**, 1063–1094 (2011)
Cheng, K.F., Serfling, R.J.: On estimation of a class of efficiency-related parameters. Scand. Actuar. J. **8**, 83–92 (1981)
Edgeworth, F.Y.: The law of error. Trans. Camb. Philos. Soc. **20**, 36–65, 113–141 (1905)
Fan, J.: On the estimation of quadratic functionals. Ann. Stat. **19**, 1273–1294 (1991)
Feller, W.: An Introduction to Probability Theory and Its Applications, vol. 2. Wiley, New York (1971)
Hallin, M., Werker, B.J.M.: Semiparametric efficiency, distribution-freeness, and invariance. Bernoulli **9**, 137–165 (2003)
Hallin, M., Oja, H., Paindaveine, D.: Semiparametrically efficient rank-based inference for shape: II Optimal R-estimation of shape. Ann. Stat. **34**, 2757–2789 (2006)
Henze, N.: On the consistency of a test for symmetry based on a runs statistic. J. Nonparametric Stat. **3**, 195–199 (1993)
Hollander, M.: Testing for symmetry. In: Johnson, N.L., Kotz, S. (eds.) Encyclopedia of Statistical Sciences, vol. 9, pp. 211–216. Wiley, New York (1988)
Jurečková, J., Sen, P.K.: Robust Statistical Procedures : Asymptotics and Interrelations. Wiley, New York (1996)
Koul, H.L.: Weighted Empirical Processes in Dynamic Nonlinear Models, 2nd edn. Springer, New York (2002)
Kraft, C.H., van Eeden, C.: Linearized rank estimates and signed-rank estimates for the general linear hypothesis. Ann. Math. Stat. **43**, 42–57 (1972)
Lehmann, E.L.: Nonparametric confidence intervals for a shift parameter. Ann. Math. Stat. **34**, 1507–1512 (1963)
Ley, C., Paindaveine, D.: Le Cam optimal tests for symmetry against Ferreira and Steel's general skewed distributions. J. Nonparametric Stat. **21**, 943–967 (2009)
McWilliams, T.P.: A distribution-free test for symmetry based on a runs statistic. J. Am. Stat. Assoc. **85**, 1130–1133 (1990)
Puri, M.L., Sen, P.K.: Nonparametric Methods in General Linear Models. Wiley, New York (1985)
Randles, R.H., Fligner, M.A., Policello, G.E., Wolfe, D.A.: An asymptotically distribution-free test for symmetry versus asymmetry. J. Am. Stat. Assoc. **75**, 168–172 (1980)
Schweder, T.: Window estimation of the asymptotic variance of rank estimators of location. Scand. J. Stat. **2**, 113–126 (1975)
Sen, P.K.: On a distribution-free method of estimating asymptotic efficiency of a class of nonparametric tests. Ann. Math. Stat. **37**, 1759–1770 (1966)

Chapter 8
Generalized MM-Tests for Symmetry

Jue Wang and David E. Tyler

Abstract Comparing the sample mean to the sample median gives rise to a test of univariate symmetry commonly referred to as the *MM*-test of symmetry. The general idea underlying this test is that the mean and median are equal at symmetric distributions, but not necessarily so at asymmetric distributions. In this paper, we study properties of a more general form of the *MM*-test of symmetry based upon the comparison of any two location estimators, and in particular upon the comparison of two different *M*-estimators of location. For these generalized *MM*-tests of symmetry, the asymptotic null distribution is obtained as well as the asymptotic distribution under local alternatives to symmetry. The local power functions help provide guidelines for choosing good members within the class of *MM*-tests of symmetry, e.g. in choosing good tuning constants for the *M*-estimators of location. A study of the local power also shows the advantages of the *MM*-test relative to other tests of symmetry. The results of the paper are shown to readily extend to testing the symmetry of the error term in a linear model.

Keywords Contiguity • Local power • M-estimates of location • Mixture models • Skew-symmetry

8.1 Introduction

In studying the properties of robust estimators of univariate location, a common assumption is that the data arises from a symmetric distribution. If the data arises from an asymmetric distribution, then different location statistics are estimating different population parameters. For example, the sample mean and the sample median estimate the population mean and population median respectively, which under asymmetric distributions are not necessarily the same. In this paper, we consider various local alternatives to symmetric distributions, including local mixture models and local skew symmetric models. Under these local alternatives,

J. Wang • D.E. Tyler (✉)
Department of Statistics and Biostatistics, Rutgers – The State University of New Jersey, New Brunswick, NJ, USA
e-mail: dtyler@rci.rutgers.edu

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_8

we first give the asymptotic distributions of robust location statistics in general, and the M-estimators of location in particular.

The main focus of this paper, though, is on the behavior of the difference between two location statistics. The standardized difference between two location statistics can be viewed as a general measure of skewness as well as a test statistic for testing the symmetry of the underlying distribution. The concept of skewness and tests for symmetry have long histories in the statistical literature, beginning with different measures on how to quantify skewness. Some of the earliest proposed measures of skewness are $(\mu - M)/\sigma$ (Pearson 1895), $(Q_{0.75} + Q_{0.25} - 2Q_{0.5})/(Q_{0.75} - Q_{0.25})$ (Bowley 1901), the classical measure of skewness $E[(X - \mu)^3]/\sigma^3$ (Edgeworth 1904; Charlier 1905), and $(\mu - Q_{0.5})/\sigma$ (Yule 1911). Here, the notation μ, M, and σ^2 represent the mean, mode, and variance, respectively, of a random variable $X \sim F$, and Q_p represents the $(p * 100)$th percentile of F. Note that except for the classical measure of skewness, these early measures of skewness correspond to a standardized difference between two location parameters. Other proposed measures of skewness also have this interpretation. For example, for $p \in (0, 1/2)$, David and Johnson (1954) suggested the following generalization of Bowley's coefficient $\mathscr{B}_p(F) = (Q_{1-p} + Q_p - 2Q_{0.5})/(Q_{1-p} - Q_p)$.

Bowley's coefficient and its generalization $\mathscr{B}_p(F)$ represent standardized differences between two different L-estimators of location, namely $(Q_{1-p} + Q_p)/2$ and the median $Q_{0.5}$, whereas Yule's coefficient can be viewed as the standardized difference between the two extremes of the Huber-type M-estimators, namely the mean and the median. A test for symmetry based on Yule's coefficient is sometimes referred to as the MM-test or mean minus the median test for symmetry, see, e.g., Hettmansperger et al. (2002). Since the mean and median may not necessarily be the two best M-estimators to compare in trying to detect asymmetry, our main goal in this paper is to generalize this test to those based upon the comparison of any two M-estimators of location. We refer to these tests as generalized MM-tests for symmetry and subsequently derive their asymptotic distributions under symmetry, as well as their local power function under the local alternatives to symmetry. The local power function provides a guideline for choosing good members within a proposed class of generalized MM-tests. For example, in using two different Huber-type M-estimators of location based on different tuning constants, one can find values for the two tuning constant which have good local power against a broad class of asymmetric distributions.

Tests based on the comparison of different M-estimators of location can be expressed in a number of asymptotically equivalent forms. For example, they are shown to be asymptotically equivalent under both the null and the local alternatives to tests based on evaluating the objective function of one of the M-estimators at the value of the other M-estimator, as well as to tests based on evaluating the M-estimating equation of one of the M-estimators at the value of the other. It is also shown that the test statistic based on the difference between a one step reweighted M-estimator of location and the initial estimator of location is asymptotically equivalent to the test statistic based on the difference between the fully iterated M-estimator and the initial estimator, even though it is well known that the one

step reweighted M-estimator is not asymptotically equivalent to the fully iterated M-estimator.

Aside from comparing the proposed tests with each other, we also compare the tests to a test for symmetry proposed by Hettmansperger et al. (2002, 2003), which is essentially based on the comparison of two different R-estimators of location. We choose to make this comparison since Hettmansperger et al. (2002, 2003) compare their test to other tests of symmetry and show that it had a superior performance in the comparisons. Due to the ability to easily tune the M-estimators, generalized MM-tests can be constructed which significantly outperform the other tests for symmetry.

The paper is organized as follows. Section 8.2 presents a brief review of the M-estimators of location and their asymptotic distributions under symmetry. In Sect. 8.3 we introduce the classes of local alternatives considered in this paper and derive the asymptotic distributions of the M-estimators under these locally asymmetric distributions. The generalized MM-tests for symmetry are introduced in Sect. 8.4 and their asymptotic distributions are derived under the null and the local alternatives. A comparison of the local power functions within classes of generalized MM-tests is given in Sect. 8.5. This allows us to make recommendation for the proper tuning of the M-estimators used in the tests. In Sect. 8.6 we compare the local power of the generalized MM-tests to the test proposed in Hettmansperger et al. (2002, 2003). Some extensions for testing symmetry in the regression setting are discussed in Sect. 8.7.

8.2 M-Estimators of Location

The concept of M-estimation was first introduced by Huber (1964) as a generalization of maximum likelihood estimation. In particular, the M-estimators of location are generalization of the maximum likelihood estimators for the center of a symmetric distribution. For a sample of size n, say $X_1, \ldots, X_n$, one way to define an M-estimator of location is as the value of μ which minimizes the objective function $\sum_{i=1}^{n} \rho(X_i - \mu)$, where $\rho(r)$ is an even non-negative function which is non-decreasing in $|r|$. If the sample represents a random sample from a continuous distribution with density $f(x - \mu)$, where f is unimodal and symmetric, then the maximum likelihood estimator of μ corresponds to choosing $\rho(r) = -\log\{f(r)\}$. Another way to define an M-estimator of location is as a solution to the M-estimating equation $\sum_{i=1}^{n} \psi(X_i - \mu) = 0$, with ψ being an odd function. If $\psi(r) = \rho'(r)$, then a root of the M-estimating equation corresponds to a critical point of the corresponding objective function.

One drawback to the above definition for the M-estimators of location is that although the resulting statistics are translation equivariant, as well as equivariant under multiplication of the data by -1, they are not necessarily scale equivariant. Consequently, one usually considers the M-estimators of location with auxiliary scale, see, e.g., Huber and Ronchetti (2009). These are generally defined either via

an objective function, i.e.

$$T_M(F_n) = \arg\max_{\mu} \sum_{i=1}^{n} \rho\left(\frac{X_i - \mu}{S(F_n)}\right), \tag{8.1}$$

where $\rho(r)$ is an even function, or as a solution $T_M(F_n)$ to the M-estimating equation

$$\sum_{i=1}^{n} \psi\left(\frac{X_i - T_M(F_n)}{S(F_n)}\right) = 0, \tag{8.2}$$

where $\psi(r)$ is an odd function. Here, F_n refers to the empirical distribution of $X_1, \ldots, X_n$, and $S(F_n)$ is an auxiliary scale statistic. An alternative expression for the M-estimating equation for location is an adaptively weighted mean, i.e.

$$T_M(F_n) = \frac{\sum_{i=1}^{n} u(R_i) X_i}{\sum_{i=1}^{n} u(R_i)}, \tag{8.3}$$

with $R_i = \{X_i - T(F_n)\}/S(F_n)$ and the weight function $u(r)$ satisfying $\psi(r) \propto ru(r)$.

By a scale statistic, we mean any non-negative statistics which is translation invariant and scale equivariant. The resulting M-estimator is then translation and scale equivariant. That is, if we let $x_i^* = a\, x_i + b$ for $i = 1, \ldots, n$ and let F_n^* denote the empirical distribution of $x_1^*, \ldots, x_n^*$, then $S(F_n^*) = |a| S(F_n)$ and $T_M(F_n^*) = a\, T_M(F_n) + b$. Hereafter, by a location statistic in general we mean any statistics $T(F_n)$ which is translation and scale equivariant.

The most common choice for a robust scale statistic is $S(F_n) = 1.4826 \cdot \text{MAD}$, where MAD refers to the median absolute deviation from the median. The factor 1.4826 is included so that $S(F_n)$ is a consistent estimator of the standard deviation when random sampling from a normal distribution. In this paper, $S(F_n)$ is taken to be any scale statistic, which may be obtained either prior to obtaining $T_M(F_n)$, i.e. a preliminary scale statistic such as 1.4826 MAD, or simultaneously with $T_M(F_n)$, e.g. as in the simultaneous M-estimators of location and scale. We adopt the usual convention, though, that the scale statistic is normalized so that $S(F_n)$ is a consistent estimator of the standard deviation when random sampling from a normal distribution. The only role $S(F_n)$ plays in this paper is through its limiting value under random sampling, which, for non-normal distributions, will not necessarily correspond to the standard deviation.

Given a random sample $X_1, \ldots, X_n$ from a distribution F, a statistic $T(F_n)$ estimates in general its functional version $T(F)$. Under certain regularity conditions, the difference between $T(F_n)$ and $T(F)$ is

$$\sqrt{n}\,\{T(F_n) - T(F)\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \text{IF}(X_i; T, F) + o_p(1), \tag{8.4}$$

see, e.g., Huber and Ronchetti (2009). The influence function of T at F is defined as $\text{IF}(x; T, F) = \lim_{\epsilon \to 0+} \{T(F_\epsilon) - T(F)\}/\epsilon$, where $F_\epsilon = (1-\epsilon)F + \epsilon\delta_x$ with δ_x representing the point mass distribution at x. Since $E_F\{\text{IF}(X; T, F)\} = 0$, this yields the asymptotic normality result

$$\sqrt{n}\,\{T(F_n) - T(F)\} \longrightarrow \textit{Normal}\left(0, \gamma^2(T, F)\right) \tag{8.5}$$

in distribution, where $\gamma^2(T, F) = E_F\left\{\text{IF}^2(X; T, F)\right\}$.

Asymptotic normality (8.5) is known to hold for M-estimators of location under general regularity conditions, see, e.g., Huber and Ronchetti (2009) or Maronna et al. (2006). Suppose now that the distribution F is symmetric about some point μ and $T(F_n)$ is a location statistic, then $T(F) = \mu$ provided $T(F)$ exists. Furthermore, for an M-estimator of location, its influence function has the simple form

$$\text{IF}(x; T_M, F) = \frac{S(F)\,\psi\,\{(x-\mu)/S(F)\}}{E_F\left[\psi'\,\{(X-\mu)/S(F)\}\right]}, \tag{8.6}$$

see, e.g., Huber and Ronchetti (2009) or Maronna et al. (2006). Consequently the limiting variance of $\sqrt{n}\,\{T_M(F_n) - \mu\}$ is

$$\gamma^2(T_M, F) = \frac{S^2(F)\,E_F\left[\psi^2\,\{(X-\mu)/S(F)\}\right]}{E_F^2\left[\psi'\,\{(X-\mu)/S(F)\}\right]}. \tag{8.7}$$

Note that μ does not correspond to the mean if the first moment of F does not exist.

Solutions to the M-estimating equations for location (8.2) are commonly found using either a Newton–Raphson approach or an iterative reweighted least squares approach, see, e.g., Huber and Ronchetti (2009). Rather than doing a full iteration of these algorithms, other location statistics can be defined by simply using one step of the algorithm based upon an initial location statistic such as the sample median. This yields the one step Newton–Raphson M-estimator

$$T_{M1}(F_n) = T_o(F_n) + S(F_n)\frac{\sum_{i=1}^n \psi\,(R_{o,i})}{\sum_{i=1}^n \psi'\,(R_{o,i})}, \tag{8.8}$$

where $R_{o,i} = \{X_i - T_o(F_n)\}/S(F_n)$, with $T_o(F_n)$ being the initial location statistic. Similarly, the one step reweighted M-estimator is defined as

$$T_{W1}(F_n) = \frac{\sum_{i=1}^n u(R_{o,i})X_i}{\sum_{i=1}^n u(R_{o,i})}. \tag{8.9}$$

Under symmetry and certain regularity conditions, the one step Newton–Raphson M-estimator is known to be asymptotically equivalent to the fully iterated M-estimator $T_M(F_n)$ associated with the same ψ-function, see, e.g., Bickel (1975). That is, the influence function for T_{M1} is also given by (8.6) and the limiting variance for

$\sqrt{n}\,\{T_{M1}(F_n) - \mu\}$ is also given by (8.7). Moreover,

$$T_{M1}(F_n) - T_M(F_n) = o_p(n^{-1/2}). \tag{8.10}$$

The one step reweighted M-estimator, however, is not asymptotically equivalent to $T_M(F_n)$. Rather, the influence function and asymptotic variance associated with T_{W1} depend upon the choice of the initial location statistic $T_o(F_n)$, see, e.g., Lopuhaä (1999). In particular, the influence function of T_{W1} is a linear combination of the influence functions of T_M and T_o, namely

$$\mathrm{IF}(x; T_{W1}, F) = (1-\lambda)\,\mathrm{IF}(x; T_M, F) + \lambda\,\mathrm{IF}(x; T_o, F), \tag{8.11}$$

where $\lambda = E_F\{u'(X)X\}/E_F\{u(X)\}$. If $\psi(r) = ru(r)$ is monotonic, then $0 < \lambda < 1$, and so (8.11) represents a convex combination of the two influence functions.

8.3 Locally Asymmetric Distributions

If the distribution is not symmetric, then the influence function and the asymptotic distribution of the M-estimators are more complicated and depend upon the influence function and asymptotic distribution of the scale statistics $S(F)$, see, e.g., Huber and Ronchetti (2009) and Maronna et al. (2006). In this section, though, we note that under local alternatives to symmetry they can have a relatively simple form. The following sequences of local alternatives considered here are:

$\mathscr{M}_{a,n}$: Local Contamination: $F(x;n) = (1-\epsilon_n)\,F(x) + \epsilon_n G(x),\ \epsilon_n = \epsilon/\sqrt{n},$

$\mathscr{M}_{b,n}$: Local Mixture: $F(x;n) = (1-\epsilon)G_1(x) + \epsilon\ G_2(x-\theta_n),\ \theta_n = \theta/\sqrt{n},$

$\mathscr{M}_{c,n}$: Local Skew-Symmetry: $f(x;n) = 2f(x)G_o\{\kappa_n(x-\mu)\},\ \kappa_n = \kappa/\sqrt{n},$ and

$\mathscr{M}_{d,n}$: Local Asymmetry: $f(x;n) = f(x)I_{[x\le\mu]} + \tau_n^{-1} f\,(x/\tau_n)\,I_{[x>\mu]},\ \tau_n = 1 + \beta/\sqrt{n},$

where $F(x;n)$ and $f(x;n)$, respectively, refer to the distribution and density function of the model. The distribution F is assumed to be symmetric about some point μ, while the distribution G can be any distribution other than one which is also symmetric about μ. The distributions G_1 and G_2 are both assumed to be symmetric about some point μ, and the distribution G_o is assumed to be symmetric about zero. Finally, the density f is assumed to be symmetric about some point μ. For absolutely continuous F, G, G_1, G_2, and G_o, we let f, g, g_1, g_2, and g_o represent their respective densities. Also, the parameters are presumed to satisfy the following constraints: $0 \le \epsilon_n \le 1$, $-\infty < \theta_n < \infty$, $\kappa_n \ge 0$, and $\tau_n > 0$. If $\epsilon_n = 0$, $\theta_n = 0$, $\kappa_n = 0$, or $\tau_n = 1$, then the corresponding model is symmetric about μ.

$\mathscr{M}_{a,n}$ is commonly used in evaluating the local power for tests of normality or tests of symmetry, see, e.g., Hettmansperger et al. (2003) and Kankainen et al. (2007). $\mathscr{M}_{b,n}$ differs from $\mathscr{M}_{a,n}$ in that the contamination proportion does not go to zero as $n \to \infty$. Rather, the contaminated distribution $F(x;n)$ approaches the symmetric mixture distribution $F(x) = (1-\epsilon)G_1(x) + \epsilon G_2(x)$, which for the special case $G_1 = G_2$ gives $F(x) = G_1(x)$. $\mathscr{M}_{c,n}$ and $\mathscr{M}_{d,n}$ correspond to other ways of modeling asymmetry. Although the exact definition of skew-symmetric distributions has yet to be unified (see Genton 2004), the class of skew-symmetric distributions used in $\mathscr{M}_{c,n}$ correspond to some of the earlier proposed forms (see Azzalini 1985). The asymmetric $\mathscr{M}_{d,n}$ was used by Gupta (1967) to study the power of a test of symmetry based on the classical skewness measure. In addition, Cassart et al. (2008) have developed optimal tests for detecting asymmetry against the local model $\mathscr{M}_{d,n}$, and they note that this model dates back to at least Fechner (1897). For any of the above models, letting $n \to \infty$ yields a distribution F which is symmetric about μ.

Let $P_{1,n}$ denote the joint distribution of $X_1, \ldots, X_n$ when they are i.i.d. X, with X having a distribution which follows one of the four local asymmetric models $\mathscr{M}_{a,n}$, $\mathscr{M}_{b,n}$, $\mathscr{M}_{c,n}$, or $\mathscr{M}_{d,n}$. Also, let $P_{o,n}$ denote this joint distribution when $\epsilon_n = 0$, $\theta_n = 0$, $\kappa_n = 0$, and $\tau_n = 1$, respectively. Thus, $P_{o,n}$ corresponds to the joint distribution of a random sample from a distribution which is symmetric about μ, namely F. The asymptotic distribution of a location statistic under the sequence $P_{o,n}$ has already been discussed in Sect. 8.2. Its asymptotic distribution under the sequence $P_{1,n}$ follows from Theorem 8.1 below. The theorem applies to any statistic $T(F_n)$ and not simply to location statistics. The proof of the theorem and the preceding lemma follow from applications of Le Cam's well-known lemmas on contiguity, see, e.g., Hájek et al. (1967), and are omitted here. Proofs for the different cases can be found in Wang (2008).

Lemma 8.1 *$P_{1,n}$ is contiguous to $P_{o,n}$ under $\mathscr{M}_{a,n}$, under $\mathscr{M}_{b,n}$ provided g_2 has a continuous second derivative, under $\mathscr{M}_{c,n}$ provided g is continuously differentiable and $g_o'(0) = 0$, and under $\mathscr{M}_{d,n}$ provided f has a continuous second derivative.*

Theorem 8.1 *If under $P_{o,n}$ the statistic $T(F_n)$ satisfies condition (8.4), and the local alternatives $\mathscr{M}_{a,n}$, $\mathscr{M}_{b,n}$, $\mathscr{M}_{c,n}$, and $\mathscr{M}_{d,n}$ satisfy the corresponding conditions in Lemma 8.1, then under the sequence $P_{1,n}$,*

$$\sqrt{n}\{T(F_n) - T(F)\} \longrightarrow \textit{Normal}\left(\eta(T,F), \gamma^2(T,F)\right)$$

in distribution, where $\gamma^2(T,F)$ is defined as in (8.5). The value of $\eta(T,F)$ depends on the particular model $\mathscr{M}_{a,n}$, $\mathscr{M}_{b,n}$, $\mathscr{M}_{c,n}$, or $\mathscr{M}_{d,n}$. These are, respectively,

$$\begin{aligned}
\eta_a(T,F) &= \epsilon\, E_G\{IF(X;T,F)\},\\
\eta_b(T,F) &= -\epsilon\theta \int_{-\infty}^{\infty} IF(x;T,F)\, g_2'(x)\, dx,\\
\eta_c(T,F) &= 2\kappa g_o(0)\, E_F\{X \cdot IF(X;T,F)\},\ \textit{and}\\
\eta_d(T,F) &= -\beta\, E_F\left(IF(X;T,F)\,\{1 + (X-\mu)f'(X)/f(X)\}\, I_{[X>\mu]}\right).
\end{aligned}$$

Using integration by parts, it can be shown that for the special case $G_1 = G_2$ in model $\mathscr{M}_{b,n}$, one always obtain $\eta_B(T, F) = \epsilon\theta$.

For location statistics, the values of $\gamma^2(T, F)$ and $\eta(T, F)$ in Theorem 8.1 do not depend upon μ. The values of the asymptotic local biases for the M-estimators of location $T_M(F_n)$ follow by using (8.6) to evaluate the expressions given in Theorem 8.1. Letting $Z = (X - \mu)/S(F)$, this gives

$$\eta_a(T_M, F) = \epsilon S(F)\, E_G\{\psi(Z)\}/E_F\{\psi'(Z)\},$$

$$\eta_b(T_M, F) = \epsilon\theta\, E_{G_2}\{\psi'(Z)\}/E_F\{\psi'(Z)\},$$

$$\eta_c(T_M, F) = 2\kappa g_o(0)\, E_F\{Z \cdot \psi(Z)\}/E_F\{\psi'(Z)\}, \text{ and}$$

$$\eta_d(T_M, F) = -\beta S(F) E_F\left(\psi(Z)\{1 + S(F)Zf'(X)/f(X)\}\, \mathrm{I}_{[Z>0]}\right)/E_F\{\psi'(Z)\}.$$

A one step Newton–Raphson M-estimator of location $T_{M1}(F_n)$ has the same value of $\gamma^2(T, F)$, $\eta_a(T, F)$, $\eta_b(T, F)$, $\eta_c(T, F)$, and $\eta_d(T, F)$ as the corresponding M-estimator $T_M(F_n)$.

8.4 Generalized *MM*-Tests for Symmetry

If two different location statistics $T_1(F_n)$ and $T_2(F_n)$ both possess the expansion (8.4), then their difference $T_D(F_n) = T_1(F_n) - T_2(F_n)$ also possesses this expansion with

$$IF(x; T_D, F) = IF(x; T_1, F) - IF(x; T_2, F).$$

Consequently, the asymptotic normality result (8.5) also holds for T_D. We presume that $\gamma^2(T_D, F) > 0$. Since $T_1(F) = T_2(F)$ for any symmetric distribution F, the rejection region for a general asymptotic α-level test for symmetry is given by

$$\chi^2(T_1, T_2) = \frac{n\, T_D^2(F_n)}{\hat{\gamma}^{\,2}(T_D, F)} > \chi^2_{1,\alpha}, \tag{8.12}$$

where $\chi^2_{1,\alpha}$ is the $1 - \alpha$ quantile of a χ^2_1 distribution, and $\hat{\gamma}^{\,2}(T_D, F)$ is some weakly consistent estimator of $\gamma^2(T_D, F)$. Note that neither the exact distribution of $T_D(F_n)$ under symmetry nor the value $\gamma^2(T_D, F)$ depend on the center of symmetric μ.

Since Theorem 8.1 applies to any statistic, it can be applied to T_D to obtain local power functions for the test (8.12). For any two location statistics satisfying (8.4), under the sequence $P_{1,n}$ for any of the models $\mathscr{M}_{a,n}$, $\mathscr{M}_{b,n}$, $\mathscr{M}_{c,n}$, or $\mathscr{M}_{d,n}$, we have

$$\chi^2(T_1, T_2) \longrightarrow \chi^2_1\left\{\delta^2(T_D, F)\right\}, \tag{8.13}$$

in distribution, where $\delta(T_D, F) = \eta(T_D, F)/\gamma(T_D, F)$. Here $\chi_1^2(\delta^2)$ refers to a non-central χ_1 distribution with non-centrality parameter δ^2. Note that $\eta(T_D, F) = \eta(T_1, F) - \eta(T_2, F)$. The expressions for η_a, η_b, η_c, and η_d are given in Theorem 8.1.

The test (8.12) gives an asymptotic α-level test over the class of symmetric distributions F for which (8.5) holds for T_D and $\hat{\gamma}^2(T_D, F)$ consistently estimates $\gamma^2(T_D, F)$. For example, consider the mean versus the median, i.e. when $T_1(F) = E_F(X)$ and $T_2(F) = F^{-1}(1/2)$. For this case, F must possess second moments, and $f(\mu) > 0$. The corresponding value of $\gamma^2(T_D, F) = \mathrm{Var}_F(X) + 1/\{2f(\mu)\}^2 - E_F(|X - \mu|)/f(\mu)$, see, e.g., Hettmansperger et al. (2003). This can be consistently estimated by estimating $\mathrm{Var}_F(X)$ by the sample variance s^2, μ by the sample median, and $E_F(|X - \mu|)$ by $\sum_{i=1}^n |X_i - T_2(F_n)|/n$, along with using a consistent estimator of $f(\mu)$.

More generally, if we take $T_{1,M}(F_n)$ and $T_{2,M}(F_n)$ to be two different M-estimators of location, then

$$\gamma^2(T_D, F) = \sum_{j=1}^{2} \frac{S_j^2(F)\, E_F\{\psi_j^2(Z_j)\}}{E_F^2\{\psi_j'(Z_j)\}} - \frac{2S_1(F)S_2(F)\, E_F\{\psi_1(Z_1)\psi_2(Z_2)\}}{E_F\{\psi_1'(Z_1)\}E_F\{\psi_2'(Z_2)\}},$$

where ψ_1 and ψ_2 are the ψ-functions associated with T_1 and T_2, respectively, S_1 and S_2 are the corresponding scale functionals, and $Z_j = (X - \mu)/S_j(F)$ for $j = 1, 2$. For smooth ψ-functions, the value of $\gamma^2(T_D, F)$ can be consistently estimated by replacing $S_1(F)$ and $S_2(F)$ with $S_1(F_n)$ and $S_2(F_n)$, respectively, replacing μ with any consistent estimator of the center of symmetry, say either $T_{1,M}(F_n)$ or $T_{2,M}(F_n)$ or their average, and by replacing the terms involving E_F with their empirical averages. For example, one can consistently estimate $E_F\{\psi_1(Z_1)\psi_2(Z_2)\}$ by

$$\frac{1}{n}\sum_{i=1}^{n} \psi_1\left(\frac{X_i - T_1(F_n)}{S_1(F_n)}\right)\psi_2\left(\frac{X_i - T_2(F_n)}{S_2(F_n)}\right).$$

For smooth and bounded ψ-functions, the corresponding asymptotic test (8.12) is valid for any absolutely continuous and symmetric F for which $T_1(F)$ and $T_2(F)$ are uniquely defined. Uniqueness always holds if in addition the ψ-functions are monotone, see, e.g., Huber and Ronchetti (2009).

The test based upon two different M-estimators of location can be expressed in a number of asymptotically equivalent forms. For example, rather than computing two M-estimators, one can use their one step Newton–Raphson versions based upon an easily computed initial location statistic. It follows from (8.10) that under either $P_{o,n}$ or $P_{1,n}$, $\chi^2(T_{1,M}, T_{2,M}) \overset{a}{\sim} \chi^2(T_{1,M1}, T_{2,M1})$. Here $Y_n \overset{a}{\sim} W_n$ means $Y_n - W_n \to 0$ in probability. Also, although the one step reweighted M-estimator is not asymptotically equivalent to the fully iterated M-estimator, if $T_o = T_{1,M}$ is used as the initial location statistic in the reweighted estimator $T_{2,W1}$, then it follows from (8.11) that $\chi^2(T_{1,M}, T_{2,M}) \overset{a}{\sim} \chi^2(T_{1,M}, T_{2,W1})$ under $P_{o,n}$ or $P_{1,n}$.

Another intuitive approach for detecting asymmetry based upon the idea of M-estimation is to consider the value of the objective function associated with one of the M-estimators when evaluated at the other M-estimator, that is to consider

$$D_1(T_{1,M}, T_{2,M}) = \sum_{i=1}^{n} \rho_1 \left(\frac{X_i - T_{2,M}(F_n)}{S_1(F_n)} \right) - \sum_{i=1}^{n} \rho_1 \left(\frac{X_i - T_{1,M}(F_n)}{S_1(F_n)} \right).$$

Note, by definition (8.1), the second summation is the minimum possible value and so $D_1(T_{1,M}, T_{2,M}) \geq 0$. A similar approach based upon plugging the value of an R-estimator of location into an objective function which yields a signed-rank estimator of location has been studied by Hettmansperger et al. (2002, 2003). Rather than consider the objective function, one could also consider the value of the M-estimating equation evaluated at some other M-estimator of location, i.e.

$$\Psi_1(T_{2,M}) = \sum_{i=1}^{n} \psi_1 \left(\frac{X_i - T_{2,M}(F_n)}{S_1(F_n)} \right).$$

By definition (8.2), $\Psi_1(T_{1,M}) = 0$. It turns out though that tests based upon $D_1(T_{1,M}, T_{2,M})$ or $\Psi_1(T_{2,M})$ are asymptotically equivalent to $\chi^2(T_{1,M}, T_{2,M})$. This follows from the following expansions, under symmetric F,

$$D_1(T_{1,M}, T_{2,M}) = \frac{\{T_{1,M}(F_n) - T_{2,M}(F_n)\}^2}{2S_1^2(F_n)} R_n + o_p(1), \text{ and} \tag{8.14}$$

$$\Psi_1(T_{2,M}) = \frac{\{T_{1,M}(F_n) - T_{2,M}(F_n)\}}{S_1(F_n)} R_n + o_p(1), \tag{8.15}$$

where $R_n = \sum_{i=1}^{n} \psi_1' (\{X_i - T_{1,M}(F_n)\}/S_1(F_n))$. Consequently, under either $P_{o,n}$ or $P_{1,n}$,

$$\chi^2(T_{1,M}, T_{2,M}) \overset{a}{\sim} \frac{2nS_1^2(F_n) D(T_{1,M}, T_{2,M})}{\hat{\gamma}^{\,2}(T_D, F) R_n} \overset{a}{\sim} \frac{nS_1^2(F_n) \Psi_1^2(T_{2,M})}{\hat{\gamma}^{\,2}(T_D, F) R_n^2}. \tag{8.16}$$

Remark 8.1 It is known that the information matrix for the location and skewness parameters, say μ and κ respectively, in a skewed normal distribution is singular at $\kappa = 0$ (see Azzalini 1985). Consequently, the local power of the likelihood ratio test for testing $\kappa = 0$, i.e. under the sequence $\mathscr{M}_{c,n}$ when $f(x) = \phi(x - \mu)$ and $G_o = \Phi$ with ϕ and Φ being the density and distribution function of the standard normal distribution respectively, is the same as the nominal significance level. This is also true for the test of symmetry given by (8.12) when using two different M-estimators of location, since, in this case, the non-centrality parameter in (8.13) equals zero.

This can be shown by first observing that under normality

$$E_F\{\psi'(Z)\} = \frac{\partial}{\partial a}\int_{-\infty}^{\infty}\psi(z+a)\phi(z)\,dz\,\Big|_{a=0} = \frac{\partial}{\partial a}\int_{-\infty}^{\infty}\psi(z)\phi(z-a)\,dz\,\Big|_{a=0}$$
$$= \int_{-\infty}^{\infty}(z-a)\psi(z)\phi(z-a)\,dz\,\Big|_{a=0}. = E_F\{Z\cdot\psi(Z)\} \tag{8.17}$$

Hence, $\eta_c(T_M, F) = 2\kappa g_o(0)$, which does not depend on the particular choice of the M-estimator, and consequently $\delta^2(T_D, F) = 0$. We conjecture that, for this case, $\eta_c(T, F) = 2\kappa g_o(0)$ for any location function, but have not formally proven this. Note that the result $\delta^2(T_D, F) = 0$ does not depend on $G_o = \Phi$. For other skewed distributions, though, i.e. when F is not normal, the non-centrality parameter is not zero in general, since (8.17) is specific to the normal distribution.

8.5 Tuning the Generalized *MM*-Tests

One advantage of using M-estimators of location over other location estimators is that they can be easily tuned while maintaining a breakdown point of $1/2$. This is not the case for L-estimates and R-estimates of location. To demonstrate the tuning of the generalized MM-tests for symmetry, we consider the behavior of the tests based on the comparison of two Huber M-estimates of location with two different tuning constants under the local mixture model $\mathscr{M}_{b,n}$. The Huber M-estimates correspond to letting ψ in (8.2) be

$$\psi_k(r) \propto \max\{-1,\ \min(r/k, 1)\}, \tag{8.18}$$

where $k > 0$ is a tuning constant. As $k \to 0$, Huber's M-estimator approaches the sample median, and as $k \to \infty$ it approaches the sample mean. If we let $G_1 = G_2$ in model $\mathscr{M}_{b,n}$, then, as noted after Theorem 8.1, the value of $\delta(T_D; F)$ in (8.13) is always zero. Consequently, we consider the case where G_2 represents a scaled version of G_1, i.e. $G_2(x) = G_1(x/b)$. Figure 8.1 shows the limiting asymptotic efficiencies as $b \to 1$ of the Huber MM-tests, using tuning constants k_1 and k_2, relative to the mean versus the median test. Here, the asymptotic relative efficiencies correspond to the ratios of the corresponding values of $\delta^2(T_D; F)$ given in (8.13). The left figure corresponds to a normal location mixture with G_1 being $Normal(\mu, \sigma^2)$, while the right figure corresponds to a mixture of location-scale t_3 distributions with G_1 now having a $t_3(\mu, \sigma)$ distribution. The asymptotic relative efficiencies do not depend on μ, σ or θ.

The axes in Fig. 8.1 correspond to $c1 = \{S_1(F)/\sigma\}k_1$ and $c2 = \{S_2(F)/\sigma\}k_2$ with $S_1(F)$ and $S_2(F)$ being the scale functionals used in the corresponding definitions of the location M-estimators. The case $c1 = c2$ is excluded since it corresponds to $\gamma^2(T_D, F) = 0$. In particular, if $S_1(F) = S_2(F)$ and $k_1 = k_2$, the two M-estimators of location are the same. One might anticipate that the relative efficiency of the

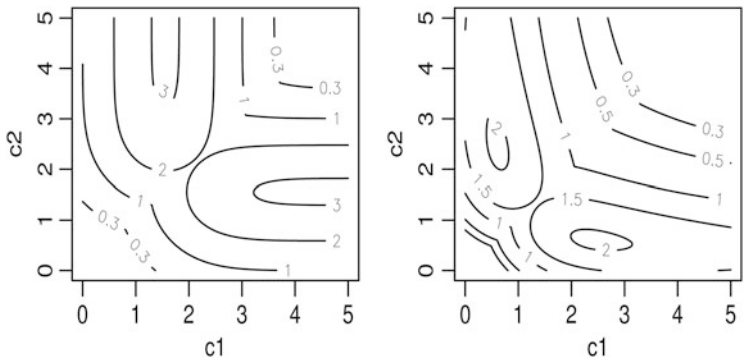

Fig. 8.1 Asymptotic efficiency under the local mixture model $\mathscr{M}_{b,n}$ of the generalized *MM*-test based on the difference of two Huber location estimators relative to the test based on the difference between the mean and the median, for the normal distribution (*left*) and the t_3 distribution (*right*)

generalized *MM*-test for symmetry would approach zero as the two estimators used in the test approach each other. However, this is not the case since it can be noted that the limiting values along the line $c1 = c2$ are positive. It is interesting to note that even when $c1$ and $c2$ are extremely close to each other, for example when both are approximately two, the asymptotic efficiency of the corresponding generalized *MM*-test can be better than the mean versus median test. This phenomena is discussed further at the end of Sect. 8.6.

When F is normal, then by convention we have $S_1(F)/\sigma = S_2(F)/\sigma = 1$, and so $c1 = k_1$ and $c2 = k_2$. From Fig. 8.1, it can be noted that for the normal mixture model, the most powerful tests in this class correspond to choosing $k_1 = \infty$, the mean, and k_2 to be approximately 1.5 rather than $k_2 = 0$, the median. Such a choice gives a test which is more than three times as efficient as the mean versus the median test. For the longer tailed t_3 mixture model, rather than using the mean and median, it is best to choose $c1$ to be approximately 2.5 and $c2$ to be approximately 0.75. The corresponding values of k_1 and k_2 then depend on the scale functional used. If we take $S_1 = S_2 = S$ to be the standard deviation, then $S(F)/\sigma = 1.732$, whereas if we choose $S_1 = S_2 = S$ to be $1.4826 \cdot \text{MAD}$, then $S(F)/\sigma = 1.134$. When using the latter scale, choosing $k_1 = 2.5$ and $k_3 = 1$ gives a good asymptotic efficiency at either the normal or the t_3 mixture models.

8.6 Comparisons with Other Tests of Symmetry

As previously noted, other robust procedures for testing symmetry have been proposed by Hettmansperger et al. (2002, 2003). In particular, within the context of testing the symmetry of the residual distribution in a linear model, they proposed a test based on the discrepancy of two robust rank-based fits, one being appropriate

for symmetric errors and the other for either symmetric and asymmetric errors. This test, denoted $H_{C\chi^2}$, can also be applied in testing univariate symmetry. In Monte Carlo studies, Hettmansperger et al. (2002) show that this test outperformed many other tests of univariate symmetry, including some robust tests of symmetry proposed by Eubank et al. (1987, 1992). Furthermore, Hettmansperger et al. (2003) show the limiting distribution of $H_{C\chi^2}$ under the local contamination model $\mathscr{M}_{a,n}$ is a non-central chi-square on one degree of freedom, and they give numerical results for the asymptotic efficiency of the $H_{C\chi^2}$-test relative to the mean minus the median test. The numerical results are under $\mathscr{M}_{a,n}$ with $F(x) = \Phi(x)$, the distribution function of a standard normal random variable, and with $G(x) = \varsigma\Phi(x-R) + (1-\varsigma)\Phi(x-L)$. For $\epsilon = 0.2$, $\varsigma = 0.95$, and $R = 1$, the asymptotic relative efficiencies are reported to be 0.9, 2.12, and $1.3 \cdot 10^5$, respectively, for $L = -2, -3$, and -4; see Table 1 of Hettmansperger et al. (2003).

Here we extend the aforementioned numerical results to obtain a comparison of the generalized *MM*-tests of symmetry to the $H_{C\chi^2}$-test. Again we consider two different Huber *M*-estimators of location for the generalized *MM*-test with tuning constants k_1 and k_2. Figure 8.2 shows the asymptotic efficiencies of the generalized *MM*-tests relative to the $H_{C\chi^2}$-test for the cases with $L = -2, -3$, and -4 as described above. The asymptotic relative efficiency corresponds to the ratio of the non-centrality parameter associated with the generalized *MM*-test to the non-centrality parameters associated with the $H_{C\chi^2}$-test. The axes c_1 and c_2 are defined as in Fig. 8.1, and since $F(x) = \Phi(x)$, we again have $c1 = k_1$ and $c2 = k_2$. If we select $k_1 = 2.5$ and $k_2 = 1$, as suggested in the previous section, the corresponding *MM*-test is locally more powerful than the $H_{C\chi^2}$-test for all three cases $L = -2, -3$, and -4 with the asymptotic relative efficiencies being 2.655, 1.969, and 1.575, respectively.

Note that when both k_1 and k_2 are large the asymptotic relative efficiencies are extremely large for the cases $L = -3$ and $L = -4$, e.g. when $k_1 = 5$ and $k_2 = 4$ the respective values are 217.7 and 5,388.2. Choosing $k_1 = 5$ and $k_2 = 4$, though, is not recommended since for the case $L = -2$ the corresponding value is only 0.120.

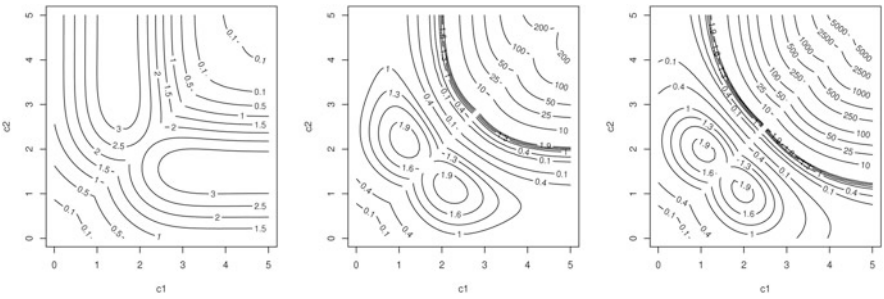

Fig. 8.2 Asymptotic relative efficiencies of the generalized *MM*-test based on the difference of two Huber location estimators relative to the $H_{C\chi^2}$-test under local contamination models $\mathscr{M}_{a,n}$, when $L = -2$ (*left*), $L = -3$ (*middle*) and $L = -4$ (*right*)

Also, such a choice has poor asymptotic efficiencies under the settings considered in Fig. 8.1. Furthermore, there is little difference between two Huber M-estimators of location with tuning constants $k_1 = 5$ and $k_2 = 4$, and so in practice round-off error may have a strong affect on the resulting MM-test. For this case, the asymptotic variance of the difference is quite small, namely $\gamma^2(T_D, \Phi) = 5.925 \cdot 10^{-6}$, whereas the squared bias terms $\eta(T_D, \Phi)$ are relatively large when $L = -3$ and $L = -4$, namely $4.721 \cdot 10^{-7}$ and $9.577 \cdot 10^{-6}$ respectively.

8.7 Extensions

Multiple Location Estimators None of the MM-tests for symmetry by themselves are able to detect all forms of asymmetry since it is always possible for two different location functionals to be equal at some asymmetric distributions. The concept behind the MM-tests can be extended to the comparison of $q > 2$ different location estimators. Here, one can generate an asymptotic χ^2_{q-1} test using the asymptotic multivariate normality of a non-singular set of $q - 1$ contrasts of the q location estimators. Again, though, it is also possible for k different location functionals to be equal at some asymmetric distribution. Alternatively, an entire class of location estimators can be used to develop graphical methods for detecting asymmetry. For example, for a univariate data set, one can plot as a function of its tuning parameter the entire Huber class of location M-estimators of location. The class of Huber location M-functionals of location, and some other classes have been shown in Wang and Tyler (2007) to represent "symmetry identifying transformations," i.e. the functionals are all equal if and only if the distribution is symmetric. Observing deviations from a flat line may be visually easier than attempting to observed asymmetric in a density estimator. Formal asymptotic tests of symmetry based on the deviation of such a plot from a flat line could also be constructed, and this would give a universal test against asymmetry. This topic is left for future research.

Regression Estimators One advantage of using M-estimators is that they readily generalize to other settings, and in particular to the regression setting. Consider the usual regression model $Y_i = \mu + \beta^T Z_i + \epsilon_i$, $i = 1, \ldots, n$, with μ and β representing the intercept and slope parameters respectively, the carriers Z_i, $i = 1, \ldots, n$ being a random sample from some non-singular p-dimensional distribution, say F_Z, and the error terms ϵ_i, $i = 1, \ldots, n$, being a random sample from the error distribution F_ϵ, with $e_i, Z_i, i = 1, \ldots, n$ being all mutually independent. The M-estimating equations for (μ, β) can be expressed as

$$\sum_{i=1}^{n} \psi\left(\frac{X_i - \hat{\mu}}{S_n}\right) = 0 \quad \text{and} \quad \sum_{i=1}^{n} \psi\left(\frac{X_i - \hat{\mu}}{S_n}\right) Z_i = 0, \tag{8.19}$$

where again $\psi(r)$ is an odd function, and $X_i = Y_i - \hat{\beta}^T Z_i$. Also, S_n is an auxiliary estimate of residual scale such that $S_n \to S(F_\epsilon)$ in probability, with $S(F_\epsilon)$ being a scale functional for the error distribution F_ϵ. Under general conditions, $\hat{\beta}$ is consistent for β whether or not the distribution of the error term ϵ is symmetric about zero. The M-estimator $\hat{\mu}$, though, is not necessarily consistent for the intercept term μ if the error distribution is not symmetric about zero. Thus, as in the location case, the difference between two M-estimators for the intercept term can be used to generate a test of symmetry about zero for the error distribution. The theory for doing so is identical to the theory for the pure location case since the asymptotic distribution of $\hat{\mu}$ is the same as that obtained by using the left-hand side of (8.19) alone but with β replacing $\hat{\beta}$. Consequently, the results obtained for the M-estimators of location in the previous sections apply directly to the M-estimators of the intercept term whenever one replaces $T_M(F_n)$, $S(F_n)$ and F with $\hat{\mu}$, S_n and $F_{\mu+\epsilon}$ respectively. In particular, the asymptotic distribution given by (8.5) and (8.7), Theorem 8.1 and the expressions for η_a, η_b, η_c and η_d given at the end of Sect. 8.3, the test statistic (8.12), its asymptotic distribution (8.13) and its asymptotic equivalence given by (8.16), as well as Figs. 8.1 and 8.2, all hold.

Acknowledgements David Tyler's research was supported in part by National Science Foundation Grant DMS-1407751. The authors are extremely grateful to the reviewers for their many helpful comments. In particular, Remark 8.1 was motivated by one of their insightful comments.

References

Azzalini, A.: A class of distributions which includes the normal ones. Scand. J. Stat. **12**, 171–178 (1985)

Bickel, P.J.: One-step Huber estimates in the linear model, J. Am. Stat. Assoc. **70**, 428–434 (1975)

Bowley, A.L.: Elements of Statistics, (4th edn., 1920). Scribner's, New York (1901)

Cassart, D., Hallin, M., Paindaveine, D.: Optimal detection of Fechner asymmetry. J. Stat. Plan. Infer. **138**, 2499–2525 (2008)

Charlier, C.V.L.: Über das Fehlergesetz. Ark. Mat. Astron. Fys. **2**, 1–9 (1905)

David, F.N., Johnson, N.L.: Statistical treatment of censored data: part I. fundamental formulae. Biometrika **41**, 228–240 (1954)

Edgeworth, F.Y.:, The law of error. Trans. Camb. Philos. Soc. **20**, 36–65, 113–141 (1904)

Eubank, R.L., LaRiccia, V.N., Rosenstein, R.B.: Test statistics derived as components of pearson's Phi-Squared distance measure. J. Am. Stat. Assoc. **82**, 816–825 (1987)

Eubank, R.L., LaRiccia, V.N., Rosenstein, R.B.: Testing symmetry about an unknown median, via linear rank procedures. J. Nonparametr. Stat. **1**, 301–311 (1992)

Fechner, G.T.: Kollectivmasslehre. Engleman, Leipzig (1897)

Genton, M.G. (ed.): Skew-Elliptical Distributions and their Applications: A Journey Beyond Normality. Chapman and Hall/CRC, Boca Raton (2004)

Gupta, M.K.: An asymptotically nonparametric test of symmetry. Ann. Math. Stat. **38**, 849–866 (1967)

Hájek, J., Šidák, Z.: Theory of Rank Tests. Academic Press, New York (1967)

Hettmansperger, T.P., McKean, J.W., Sheather, S.J.: Finite sample performance of tests for symmetry of the errors in a linear model. J. Stat. Comput. Simul. **72**, 863–879 (2002)

Hettmansperger, T.P., McKean, J.W., Sheather, S.J.: Testing symmetry of the errors of a linear model. In: Moore, M., Froda, S., Léger, C. (eds.) Mathematical Statistics and Applications: Festschrift for Constance van Eeden. Lecture Notes - Monograph Series, pp. 99–112. Institute of Mathematical Statistics (2003)

Huber, P.J.: Robust estimation of a location parameter, Ann. Math. Stat. **35**, 73–101 (1964)

Huber, P.J., Ronchetti, E.M.: Robust Statistics, 2nd edn. Wiley, New York (2009)

Kankainen, A., Taskinen, S. Oja, H.: Tests of multinormality based on location vectors and scatter matrices, Stat. Methods Appl. **16**, 357–379 (2007)

Lopuhaä, H.P: Asymptotics of reweighted estimators of multivariate location and scatter. Ann. Stat. **27**, 1638–1665 (1999)

Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust Statistics: Theory and Methods. Wiley, Chichester (2006)

Pearson, K.: Contributions to the mathematical theory of evolution II. Skew variation in homogeneous material. Philos. Trans. R. Soci. Lond. A **186**, 343–414 (1895)

Wang, J., Tyler, D.E.: A graphical method for detecting asymmetry. In: Transactions of the 63rd Deming Conference, Dec 2007

Wang, J.: Some properties of robust statistics under asymmetric models. Ph.D. Dissertation, Department of Statistics, Rutgers–The State University of New Jersey (2008)

Yule, G.U.: Introduction to the Theory of Statistics. Griffin, London (1911)

Part III
Nonparametric and Robust Methods for Multivariate and Functional Data

Chapter 9
M-Estimators of the Correlation Coefficient for Bivariate Independent Component Distributions

Georgy Shevlyakov and Pavel Smirnov

Abstract A few historical remarks on the notion of correlation, as well as a brief review of robust estimators of the correlation coefficient are given. A family of M-estimators of the correlation coefficient for bivariate independent component distributions is proposed. Consistency and asymptotic normality of these estimators are established, and the explicit expression for their asymptotic variance is obtained. A minimax variance (in the Huber sense) M-estimator of the correlation coefficient for ε-contaminated bivariate normal distributions is designed. Although the structure of this new result generally is similar to the former minimax variance M-estimator of the correlation coefficient proposed by Shevlyakov and Vilchevski (Stat. Probab. Lett. **57**, 91–100, 2002b), the efficiency of this new estimator is considerably greater than that of the former one as it generalizes the maximum likelihood estimator of the correlation coefficient of the bivariate normal distribution. Furthermore, highly efficient and robust estimators of correlation are obtained by applying highly efficient and robust estimators of scale. Under the ε-contaminated bivariate normal, t- and independent component Cauchy distributions, the proposed robust estimators dominate over the sample correlation coefficient. The comparative analytical and Monte Carlo study of various robust estimators confirm the effectiveness of the proposed M-estimator of the correlation coefficient.

Keywords Bivariate normal distribution • Correlation coefficient • Independent component distributions • M-estimator • Robustness

G. Shevlyakov (✉) • P. Smirnov
Institute of Applied Mathematics and Mechanics, Peter the Great St. Petersburg Polytechnic University, 29 Polytechnicheskaya ul., 195251, St. Petersburg, Russia
e-mail: Georgy.Shevlyakov@phmf.spbstu.ru; psmirnov@spbstu.ru

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_9

9.1 Introduction

9.1.1 Historical Remarks on the Notion of Correlation

The term "correlation" was introduced into science by a French naturalist Georges Cuvier (1769–1832), one of the major figures in natural sciences in the early nineteenth century, who had established paleontology and comparative anatomy. The word "correlation" is of the late Latin origin meaning "association," "connection," "correspondence," "interdependence," "relationship."

Cuvier discovered and studied the relationships between the parts of animals, between the structure of animals and their mode of existence, between the species of animals and plants, and many others. This experience made him formulate the general principles of "the correlation of parts" and of "the functional correlation." From Cuvier to Galton, correlation had been understood as a qualitatively described relationship, not deterministic but of a statistical nature, however observed at that time within a rather narrow area of phenomena.

Francis Galton (1822–1911), a British anthropologist, biologist, psychologist, and meteorologist, understood that correlation is the interrelationship in average between any random variables. Correlation analysis (this term also was coined by Galton) deals with estimation of the value of correlation by number indexes or coefficients. Galton contributed much to science studying the problems of heredity of qualitative and quantitative features. They were numerically examined by Galton on the basis of the concept of correlation. Until present, the data on demography, heredity, and sociology collected by Galton with the corresponding numerical examples of correlations computed are used.

Karl Pearson (1857–1936), a British mathematician, statistician, biologist, philosopher, had written out the explicit formulas for the population correlation coefficient

$$\rho = \frac{\mathbf{C}\mathrm{ov}(X, Y)}{[\mathbf{V}\mathrm{ar}(X)\ \mathbf{V}\mathrm{ar}(Y)]^{1/2}} \tag{9.1}$$

and its sample version (Pearson 1894)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}} \tag{9.2}$$

(here $\bar{x}$ and $\bar{y}$ are the sample means of the observations $\{x_i\}$ and $\{y_i\}$ of random variables X and Y). However, Pearson did not definitely distinguish the population and sample versions of the correlation coefficient, as it is commonly done at present.

On the one hand, the sample correlation coefficient r is a statistical counterpart of the correlation coefficient ρ of a bivariate distribution, where $\mathbf{V}\mathrm{ar}(X)$, $\mathbf{V}\mathrm{ar}(Y)$, and $\mathbf{C}\mathrm{ov}(X, Y)$ are the variances and the covariance of the random variables X and

Y, respectively. On the other hand, it is an efficient maximum likelihood estimator of the correlation coefficient ρ of the bivariate normal distribution

$$N(x,y;\mu_X,\mu_Y,\sigma_X,\sigma_Y,\rho)=\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\right.$$
$$\left.\times\left[\frac{(x-\mu_X)^2}{\sigma_X^2}-2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}, \qquad (9.3)$$

where $\mu_X=E(X)$, $\mu_Y=E(Y)$, $\sigma_X^2=\mathbf{V}\mathrm{ar}(X)$, $\sigma_Y^2=\mathbf{V}\mathrm{ar}(Y)$ (Kendall and Stuart 1963).

Galton (1886) derived the bivariate normal distribution density (9.3), and he was the first who used it to scatter the frequencies of children's stature and parents' stature. Pearson noted that "in 1885 Galton had completed the theory of bi-variate normal correlation" (Pearson 1920, p. 37). Like Galton, Auguste Bravais (1846), a French naval officer and astronomer, came very near to formula (9.1) when he called one parameter of the bivariate normal distribution "une correlation," but he did not recognize it as a measure of the interrelationship between variables. However, "his work in Pearson's hands proved useful in framing formal approaches in those areas" (Stigler 1986, p. 353).

Pearson's formulas (9.1) and (9.2) proved to be very fruitful for studying dependencies: correlation analysis and most of multivariate statistical analysis tools are based on the pair-wise Pearson correlations; we may also add the correlation and spectral theories of stochastic processes, etc. Since the time Pearson introduced the sample correlation coefficient (9.2), many other measures of correlation have been used aiming at estimation of the closeness of interrelationship (the coefficients of association, determination, contingency, etc.). Some of them were proposed by Pearson (1948).

9.1.2 *A Brief Review of Robust Estimators of the Correlation Coefficient*

The high sensitivity of the sample correlation coefficient r to outliers in the data described by the mixture of normal densities (Tukey's 1960 gross error model) ($0\le\varepsilon<0.5$)

$$f(x,y)=(1-\varepsilon)N(x,y;\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)+\varepsilon N(x,y;\mu_1',\mu_1',\sigma_1',\sigma_2',\rho') \qquad (9.4)$$

and the necessity in its robust counterparts was firstly considered in Gnanadesikan and Kettenring (1972). The first and second summands in (9.4) generate "good" and "bad" data, respectively: in general, the component means μ_1' and μ_2', standard deviations σ_1' and σ_2', as well as the correlation coefficient ρ' of "bad" data may significantly differ from their counterparts in the first summand. Following

the main paradigm of robust estimation, we are most interested in estimation of the correlation coefficient ρ of "good" data regarding "bad" data as outliers. In model (9.4), the sample correlation coefficient r can be strongly biased with respect to ρ up to the change of its sign (Gnanadesikan and Kettenring 1972; Devlin et al. 1975).

At present there exist two main approaches to design of robust estimators, i.e., the minimax principle of quantitative robustness (Huber 1981), and the approach of qualitative robustness based on influence functions (Hampel et al. 1986). With the first approach, the least informative distribution minimizing Fisher information over a given class of distributions is determined followed by the subsequent use of the optimal maximum likelihood estimator for this distribution. In this case, the minimax approach ensures that the asymptotic variance of the optimal estimator will not exceed a certain threshold. According to the second approach, an estimator with the desired influence function whose type of behavior determines the qualitative robustness properties of an estimator such as its sensitivity to the presence of gross outliers in the data, to the data rounding off, etc.

Most of the earlier results on robust estimation of the correlation coefficient have been obtained from heuristic considerations related to the desired behavior of their influence functions (Gnanadesikan and Kettenring 1972; Devlin et al. 1975). Huber (1964, 1981) used the minimax approach to design the minimax variance and bias estimators of location and scale, as well as of correlation.

In our former works (Pasman and Shevlyakov 1987; Shevlyakov and Vilchevski 2002a; Shevlyakov and Smirnov 2011), a classification of robust estimators of the correlation coefficient is given. The group of robust estimators of correlation based on robust estimators of the variances of the principal components originally proposed by Gnanadesikan and Kettenring (1972) turned out to be most prospective among others: the minimax variance and bias (in Huber's sense) estimators such as the trimmed r_{TRIM} and MAD r_{MAD} correlation coefficients belong to this group (Shevlyakov and Vilchevsky 2002b; Shevlyakov et al. 2012). Moreover, these estimators are intrinsically conformed to the family of independent component distributions (ICD) (Shevlyakov and Vilchevsky 2002b; Shevlyakov et al. 2012).

In this paper we generalize the robust minimax variance M-estimators of the correlation coefficient (Shevlyakov and Vilchevsky 2002b), namely r_{TRIM} and r_{MAD}, introducing M-estimators of the correlation coefficient as the extension of the maximum likelihood estimators for the ICD family (quite similarly to Huber's (1964) program for robust estimation of location) thus enhancing the efficiency of these new estimators of correlation. This work concludes the cycle of works on robust minimax variance and bias M-estimators of correlation for the ICD family (Shevlyakov and Vilchevsky 2002b; Shevlyakov and Smirnov 2011; Shevlyakov et al. 2012).

An outline of the remainder of the paper is as follows. In Sect. 9.2, the ICD family and some former results on the minimax estimation of correlation are briefly described. In Sect. 9.3, new M-estimators of the correlation coefficient for the ICD family are proposed. These M-estimators possess the minimax variance property (in Huber's sense) in ε-contaminated bivariate normal distribution models. In Sect. 9.4,

a comparative Monte Carlo study of various robust estimators of the correlation coefficient is performed. In Sect. 9.5, some conclusions are drawn.

9.2 Bivariate Independent Component Distributions and Earlier Results on Robust Minimax Estimation

In this section, we follow Shevlyakov and Vilchevsky (2002b), Shevlyakov and Smirnov (2011), and Shevlyakov et al. (2012) and recall the family of the bivariate ICD densities and the minimax estimators based on them, as they are necessary for our further constructions.

9.2.1 Bivariate Independent Component Distribution Densities

ICD densities with unknown but equal variances (the parameters of location of the random variables X and Y are assumed known) are defined as follows (Shevlyakov and Vilchevski 2002a; Shevlyakov and Vilchevsky 2002b)

$$f(u,v) = \frac{1}{\sigma\sqrt{1+\rho}}\, g\left(\frac{u}{\sigma\sqrt{1+\rho}}\right) \frac{1}{\sigma\sqrt{1-\rho}}\, g\left(\frac{v}{\sigma\sqrt{1-\rho}}\right), \tag{9.5}$$

where σ is the standard deviation; ρ is the correlation coefficient, u and v are the principal variables $u = (x+y)/\sqrt{2}$, $v = (x-y)/\sqrt{2}$; $g(x)$ is a symmetric density belonging to a certain class $\mathcal{G}$.

The family (9.5) contains the standard bivariate normal distribution density $f(x,y) = N(x,y|0,0,1,1,\rho)$ when $g(z) = \varphi(z) = (2\pi)^{-1/2}\exp(-z^2/2)$. Further, we use the bivariate heavy-tailed Cauchy ICD density; in this case, ρ is regarded as a parameter of correlation.

The idea of the ICD densities given by (9.5) is quite plain: for any pair (X,Y), the transformation $U = X+Y$, $V = X-Y$ gives the uncorrelated random principal variables (U,V), actually independent for densities (9.5). Thus, estimation of their scales $S(U) = \sigma\sqrt{1+\rho}$ and $S(V) = \sigma\sqrt{1-\rho}$ solves the problem of estimation of the correlation coefficient between X and Y by using

$$\hat{\rho} = \frac{\hat{S}^2(U) - \hat{S}^2(V)}{\hat{S}^2(U) + \hat{S}^2(V)}, \tag{9.6}$$

since the following identity for the correlation coefficient holds (Gnanadesikan and Kettenring 1972)

$$\rho = \frac{S(U)^2 - S(V)^2}{S(U)^2 + S(V)^2}.$$

Thus, the class (9.6) of estimators entirely corresponds to the ICD family (9.5), and this allows to extend Huber's results on robust minimax M-estimators of scale to robust estimation of the correlation coefficient.

9.2.2 Minimax Variance Robust Estimation of Correlation

In Shevlyakov and Vilchevsky (2002b), it is shown that under regularity conditions the estimator $\hat{\rho}$ (9.6) is consistent and asymptotically normal with the following variance

$$AV(\hat{\rho}) = 2(1-\rho^2)^2\, V(\chi, g), \quad V(\chi, g) = \frac{\int \chi^2(x) g(x)\, dx}{\left(\int x\chi'(x) g(x)\, dx\right)^2}. \tag{9.7}$$

where $V(\chi, g)$ is the asymptotic variance of Huber's M-estimates of scale $\widehat{S}(U)$ and $\hat{S}(V)$ in formula (9.6) (Huber 1981).

The remarkable feature of formula (9.7) for the asymptotic variance is that it has two factors: the first depends only on ρ, the second $V(\chi, g)$, as it was aforementioned, is exactly the asymptotic variance of M-estimators of scale. Thus Huber's minimax variance M-estimators of scale in the gross error model (Huber 1981, pp. 120–122) can be directly applied for the minimax variance estimation of the correlation coefficient for ε-contaminated bivariate normal distributions

$$f(x, y) \geq (1-\varepsilon)\, N(x, y; 0, 0, 1, 1, \rho), \qquad 0 \leq \varepsilon < 1 \tag{9.8}$$

giving the trimmed correlation coefficient (Shevlyakov and Vilchevsky 2002b)

$$r_{TRIM} = \left(\sum_{i=n_1+1}^{n-n_2} u_{(i)}^2 - \sum_{i=n_1+1}^{n-n_2} v_{(i)}^2 \right) \Bigg/ \left(\sum_{i=n_1+1}^{n-n_2} u_{(i)}^2 + \sum_{i=n_1+1}^{n-n_2} v_{(i)}^2 \right), \tag{9.9}$$

where $u_{(i)}^2$ and $v_{(i)}^2$ are the ith order statistics of the squared robust principal variables; n_1 and n_2 are the numbers of trimmed observations. The levels of trimming n_1 and n_2 of the minimax trimmed correlation coefficient r_{TRIM} depend on the contamination parameter ε: $n_1 = n_1(\varepsilon)$ and $n_2 = n_2(\varepsilon)$ (Shevlyakov and Vilchevsky 2002b).

In particular, the minimax variance estimator r_{TRIM} takes the following limit form as $\varepsilon \to 1$: it tends to the MAD-correlation coefficient r_{MAD}

$$r_{MAD} = \frac{\mathrm{MAD}^2 u - \mathrm{MAD}^2 v}{\mathrm{MAD}^2 u + \mathrm{MAD}^2 v}, \tag{9.10}$$

where u and v are the robust principal variables

$$u = \frac{x - \operatorname{med} x}{\sqrt{2}\,\mathrm{MAD}\, x} + \frac{y - \operatorname{med} y}{\sqrt{2}\,\mathrm{MAD}\, y}, \qquad v = \frac{x - \operatorname{med} x}{\sqrt{2}\,\mathrm{MAD}\, x} - \frac{y - \operatorname{med} y}{\sqrt{2}\,\mathrm{MAD}\, y}. \tag{9.11}$$

Concluding the review of our former results, we note that the same minimax variance robust estimators of the correlation coefficient, namely, r_{TRIM} and r_{MAD}, are also minimax bias robust: this is due to the asymptotic bias of estimators (9.6) whose structure is similar to that of formula (9.7) (for details, see Shevlyakov et al. 2012).

9.3 Main Result: M-Estimators of the Correlation Coefficient and their Minimax Variance Properties

In this section, we implement a classical Huber's approach (Huber 1964) to designing robust estimators of the correlation coefficient.

9.3.1 The Maximum Likelihood Estimator of the Correlation Coefficient

Consider the maximum likelihood estimator of the correlation coefficient ρ for independent component distributions $f(u, v)$ (9.5), where we without any loss of generality set $\sigma = 1$

$$\sum_{i=1}^{n} \psi_{ML}(u_i, v_i; \hat{\rho}_{ML}) = 0 \tag{9.12}$$

with the score function $\psi_{ML}(u, v; \rho) = \partial \log f(u, v; \rho)/\partial\rho$. It has the following form

$$\psi_{ML}(u, v; \rho) = \frac{1}{1+\rho}\chi_{ML}\left(\frac{u}{\sqrt{1+\rho}}\right) - \frac{1}{1-\rho}\chi_{ML}\left(\frac{v}{\sqrt{1-\rho}}\right), \tag{9.13}$$

where χ_{ML} is the maximum likelihood score function for M-estimators of scale (Huber 1981)

$$\chi_{ML}(t) = -(1 + tg'(t)/g(t)). \tag{9.14}$$

In the particular case of the bivariate normal distribution when $g(z) = \varphi(z)$, the routine calculation of the Fisher information $I(\rho)$ for the correlation coefficient yields

$$I(\rho) = \frac{1+\rho^2}{(1-\rho^2)^2},$$

which in its turn gives the classical value of the asymptotic variance (Kendall and Stuart 1963)

$$AV(\hat{\rho}_{ML}) = \frac{1}{I(\rho)} = \frac{(1-\rho^2)^2}{(1+\rho^2)}\,. \tag{9.15}$$

From the comparison of this result with the asymptotic variance $AV(r) = (1-\rho^2)^2$ of the sample correlation coefficient r at the bivariate normal distribution it follows that the maximum likelihood estimator $\hat{\rho}_{ML}$ is asymptotically more accurate than r being equal to it only at $\rho = 0$.

9.3.2 M-Estimators of the Correlation Coefficient

Taking into account the structure of the maximum likelihood estimating Eq. (9.12) and the corresponding score function (9.13) for estimation of the correlation coefficient, we implement Huber's program realized for robust estimation of location and scale (Huber 1981) and define a class of M-estimators of ρ for independent component distributions with an arbitrary score function χ

$$\sum_{i=1}^{n} \psi_M(u_i, v_i; r_M) = 0\,, \tag{9.16}$$

where

$$\psi_M(u, v; \rho) = \frac{1}{1+\rho}\chi\left(\frac{u}{\sqrt{1+\rho}}\right) - \frac{1}{1-\rho}\chi\left(\frac{v}{\sqrt{1-\rho}}\right).$$

The particular case of an M-estimator of the correlation coefficient is given by $\chi_{MAD}(z) = \text{sign}(|z|-1)$, the score function for the MAD-estimator of scale in (9.16), with the corresponding $MMAD$-estimator r_{MMAD}.

9.3.3 Asymptotic Variance of M-Estimators

Now we show that the asymptotic variance of M-estimators has the structure similar to that of formula (9.7).

Theorem 9.1 *Assume that the following regularity conditions imposed on symmetric densities g and scores χ hold:*

($g1$) *$g(x)$ is twice continuously differentiable and satisfies $g(x) > 0$ for all x in $\mathbb{R}$.*
($g2$) *Fisher information for scale $I(g)$ satisfies $0 < I(g) < \infty$.*

($\chi 1$) *A score function* χ *is well defined and continuous on* $\mathbb{R} \setminus C(\chi)$*, where the set* $C(\chi)$ *of discontinuity points of* $\chi(\cdot)$ *is finite. In each point of* $C(\chi)$ *there exist finite left and right limits of* χ *which are different. Also* $\chi(-x) = \chi(x)$ *if* $(-x, x) \subset \mathbb{R} \setminus C(\chi)$*, and there exists* $d > 0$ *such that* $\chi(x) \le 0$ *on* $(0, d)$ *and* $\chi(x) \ge 0$ *on* (d, ∞).

($\chi 2$) *The set* $D(\chi)$ *of points in which* χ *is continuous but in which* χ' *is not defined or not continuous is finite.*

($\chi 3$) $\int \chi(x)g(x)\,dx = 0$ *and* $\int \chi^2(x)g(x)\,dx < \infty$.

($\chi 4$) $0 < \int x\chi'(x)g(x)\,dx < \infty$.

Then, in the class of bivariate independent component distributions (9.5)*, the asymptotic variance of M-estimators has the form*

$$AV(r_M) = \frac{2(1-\rho^2)^2}{1+\rho^2} V(\chi, g)\,, \tag{9.17}$$

where

$$V(\chi, g) = \frac{\int \chi^2(x)g(x)\,dx}{\left(\int x\chi'(x)g(x)\,dx\right)^2}$$

is the asymptotic variance of M-estimators of scale.

Proof Here we give the sketch of proof. Conditions ($g1$)–($\chi 4$) are sufficient for consistency and asymptotic normality of the M-estimator r_M. This result follows from the consistency and asymptotic normality of the M-estimators of scale defined by score functions χ (Shevlyakov and Vilchevsky 2002b). The asymptotic variance of r_M is obtained by direct computation from the following formula of Hampel et al. (1986)

$$AV(r_m) = V(\psi_M, f) = \frac{A}{B^2},$$

where

$$A = E(\psi_M^2) = \int\int \psi_M^2(u, v; \rho) f(u, v; \rho)\,du\,dv,$$

$$B = E(\partial\psi_M/\partial\rho) = \int\int \frac{\partial\psi_M(u, v; \rho)}{\partial\rho} f(u, v; \rho)\,du\,dv.$$

□

In this paper, we use the well-balanced set of regularity conditions proposed by Hampel et al. (1986, pp. 125, 139): the comments on those conditions can be found in Shevlyakov et al. (2012).

Formula (9.17) for the asymptotic variance of r_M is similar to formula (9.7) also having two factors: the first depends only on ρ, the second $V(\chi, g)$ is the asymptotic variance of M-estimators of scale. The difference is only in the factor $(1+\rho^2)$ in the denominator of (9.17), which reduces the variance value and, therefore, enhances the estimator efficiency.

Moreover, most results on robust estimation of scale (Huber 1981; Hampel et al. 1986) can be directly applied to robust estimation of the correlation coefficient of bivariate independent component distributions. The further result is based on Theorem 9.1 and Huber's results on minimax variance estimation of scale (Huber 1981, pp. 120–121).

9.3.4 *Minimax Variance M-Estimators of the Correlation Coefficient*

To design minimax variance M-estimators of the correlation coefficient, we recall Huber's result on minimax variance M-estimators of scale (Huber 1981): under the conditions of regularity $(g1)-(\chi 4)$ of Theorem 9.1, M-estimators $\hat{S}$ of scale defined by the estimating equation $\sum \chi(x_i/\hat{S}) = 0$ are consistent, asymptotically normal and possess the minimax (saddle-point) property with regard to the asymptotic variance $V(\chi, g) = AV(\hat{S})$ (9.7)

$$V(\chi^*, g) \le V(\chi^*, g^*) \le V(\chi, g^*). \tag{9.18}$$

Here g^* is the least informative density minimizing Fisher information $I(g)$ for scale

$$g^* = \arg\min_{g\in\mathcal{G}} I(g), \quad I(g) = \int \left(-x\,\frac{g'(x)}{g(x)} - 1\right)^2 g(x)\,dx \tag{9.19}$$

over the class of ε-contaminated normal distribution densities

$$\mathcal{G} = \{g:\ g(x) \ge (1-\varepsilon)\varphi(x)\},\ 0 \le \varepsilon < 1. \tag{9.20}$$

In this case, the optimal score function $\chi^*(x)$ is given by

$$\chi^*(x) = \begin{cases} x_0^2 - 1, & \text{for} \quad |x| < x_0, \\ x^2 - 1, & \text{for} \quad x_0 \le |x| \le x_1, \\ x_1^2 - 1, & \text{for} \quad |x| > x_1, \end{cases} \tag{9.21}$$

where the parameters $x_0(\varepsilon)$ and $x_1(\varepsilon)$ depend on the contamination parameter ε being tabulated in Huber (1981, , p. 121).

Note that the median absolute deviation $\hat{S} = \text{MAD}\ x$ is a limit case of the minimax variance estimator as $\varepsilon \to 1$ with $\chi^*(x) = \chi_{MAD}(x) = \text{sign}(|x| - 1)$.

The saddle-point inequality (9.18) shows that the estimator $\hat{S}$ determined by the score function χ^* provides the guaranteed level of the accuracy of estimation for all g in $\mathcal{G}$

$$V(\chi^*, g) \le V(\chi^*, g^*) = 1/(n\,I(g^*)).$$

The following result is obtained by the direct application of the above solution.

Theorem 9.2 *In the subclass* (9.5) *of* ε*-contaminated bivariate normal distributions*

$$f(x, y) \ge (1 - \varepsilon)\, N(x, y|0, 0, 1, 1, \rho), \qquad 0 \le \varepsilon < 1, \tag{9.22}$$

the minimax variance estimator of ρ *is given by the M-estimator* (9.16) *with the score function* (9.21)*, where* $x_1 = x_1(\gamma)$ *and* $x_2 = x_2(\gamma)$ *depend on the value of the contamination parameter* ε *through the auxiliary parameter* $\gamma = 1 - \sqrt{1 - \varepsilon}$.

The proof of Theorem 9.2 literally repeats the proof of Theorem 9.2 given in Shevlyakov and Vilchevsky (2002b).

If $\varepsilon = 0$, then the minimax variance M-estimator of the correlation coefficient coincides with the maximum likelihood estimator defined by (9.13)–(9.14). In the limit case as $\varepsilon \to 1$, the minimax variance M-estimator tends to the *MMAD*-estimator (9.16) with the score function $\chi_{MAD}(x) = \text{sign}(|x| - 1)$.

Under the aforementioned regularity conditions $(g1)$–$(\chi 4)$ imposed on score functions χ and densities g, the most V-robust estimator of the correlation coefficient in the sense of Theorem 10 in Hampel et al. (1986, , pp. 142–143) is given by the estimator (9.16) with the optimal score $\chi_{MAD}(x) = \text{sign}(|x| - 1)$.

9.3.5 Highly Robust and Efficient M-Estimators of the Correlation Coefficient

Another way to enhance robustness and efficiency of M-estimators (9.16) of the correlation coefficient is to use in their structure the score functions of highly robust and efficient estimators of scale.

Rousseeuw and Croux (1993) proposed one of the most robust and efficient estimators of scale, namely the Q_n, later became commonly used in practice. The Q_n is defined as a lower quartile of pairwise distances: $Q_n = \{|x_i - x_j|\}_k$, $k = \binom{h}{2}$, $h = [n/2] + 1$, and it has the asymptotic efficiency of 82.3 % and the breakdown point of 50 %. A serious drawback is its large computational complexity, as $O(n^2)$ pairwise differences are involved.

In Smirnov and Shevlyakov (2014), a parametric family of M-estimators of scale based on the Rousseeuw–Croux Q_n-estimator is proposed; the estimator bias and efficiency are studied. A low-complexity one-step M-estimator is obtained

allowing a considerably faster computation with the greater than 80 % efficiency at a Gaussian and the highest possible 50 % breakdown point. We denote this M-estimator approximating the Q_n-estimator as FQ_n; it has the following score function

$$\chi_{FQ}(x) = 1/\sqrt{\pi} - 2\varphi(x). \tag{9.23}$$

Substituting this score function into (9.16), we get a highly robust and efficient M-estimator of the correlation coefficient denoted as r_{MFQ}.

9.4 Monte Carlo Performance Evaluation

9.4.1 Monte Carlo Experiment Set-up

In Tables 9.1, 9.2, 9.3, and 9.4, we exhibit experimental results (50,000 trials) on the comparative performance of the proposed and classical estimators on small ($n = 20$) and large ($n = 1000$) samples at the bivariate normal and ε-spherically contaminated normal distributions with the density

$$f(x,y) = (1-\varepsilon)N(x,y;0,0,1,1,\rho) + \varepsilon N(x,y;0,0,k,k,\rho')\,,$$

the bivariate ICD Cauchy density (9.5) with $g(z) = \pi^{-1}(1+z^2)^{-1}$ and the bivariate Cauchy t-distribution (a particular case of the bivariate t-distribution) with the density

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\left(1 + \frac{x^2+y^2-2\rho xy}{1-\rho^2}\right)^{-3/2}$$

Table 9.1 Normal distribution: $\rho = 0.9$

	r	r_Q	r_S	r_K	r_{MAD}	r_{FQ}	r_{MFQ}	r_{MMAD}	r_{MCD}
$n = 20$									
Mean	0.895*	0.858	0.875	0.892	0.873	0.889	**0.897**	0.906	0.874
MSE	0.050	0.139	0.069	0.063	0.093	0.056*	**0.037**	0.074	0.154
RE	0.729*	0.103	0.431	0.467	0.224	0.596	**1.323**	0.334	0.078
$n = 1000$									
Mean	0.900	0.899	0.900	0.900	0.899	0.900	0.900	0.900	0.900
MSE	0.006*	0.015	0.007	0.007	0.010	0.007	**0.005**	0.007	0.010
RE	0.987*	0.154	0.739	0.820	0.363	0.804	**1.453**	0.662	0.380

Table 9.2 Contaminated normal distribution: $\rho = 0.9$, $\varepsilon = 0.1$, $\rho' = 0$, $k = 3$

	r	r_Q	r_S	r_K	r_{MAD}	r_{FQ}	r_{MFQ}	r_{MMAD}	r_{MCD}
$n = 20$									
Mean	0.641	0.789	0.708	0.777	0.855	0.849	0.831	0.875*	**0.876**
MSE	0.504	0.205	0.269	0.195	0.103	**0.088**	0.101*	**0.088**	0.143
RE	0.014	0.061	0.051	0.079	0.194	**0.295**	0.273*	0.256	0.091
$n = 1000$									
Mean	0.553	0.835	0.730	0.787	0.890*	0.868	0.845	0.876	**0.899**
MSE	0.452	0.059	0.172	0.105	0.015*	0.024	0.046	0.026	**0.009**
RE	0.010	0.087	0.052	0.101	0.304	**0.465**	0.418	0.414	0.438*

Table 9.3 ICD Cauchy distribution: $\rho = 0.9$

	r	r_Q	r_S	r_K	r_{MAD}	r_{FQ}	r_{MFQ}	r_{MMAD}	r_{MCD}
$n = 20$									
Mean	0.624	0.681	0.624	0.716	0.856*	**0.857**	0.667	0.768	0.823
MSE	0.625	0.312	0.352	0.272	0.148*	**0.143**	0.281	0.202	0.258
RE	0.006	0.037	0.038	0.045	0.090*	**0.097**	0.073	0.077	0.030
$n = 1000$									
Mean	0.628	0.743	0.743	0.639	**0.899**	**0.899**	0.688	0.799	**0.899**
MSE	0.613	0.160	0.159	0.263	0.014*	**0.013**	0.213	0.103	0.018
RE	0.000	0.046	0.051	0.038	0.198*	**0.212**	0.074	0.111	0.107

Table 9.4 Bivariate Cauchy t-distribution: $\rho = 0.9$

	r	r_Q	r_S	r_K	r_{MAD}	r_{FQ}	r_{MFQ}	r_{MMAD}	r_{MCD}
$n = 20$									
Mean	0.844	0.852	0.836	**0.884**	0.872	0.880*	0.679	0.777	0.876
MSE	0.295	0.145	0.140	0.096*	0.105	**0.082**	0.258	0.184	0.148
RE	0.022	0.096	0.117	0.203*	0.177	**0.288**	0.103	0.097	0.085
$n = 1000$									
Mean	0.846	0.899	0.856	**0.900**	0.899	**0.900**	0.689	0.799	**0.900**
MSE	0.289	0.015	0.047	0.011*	0.011*	**0.009**	0.212	0.103	0.012
RE	0.000	0.154	0.132	0.306*	0.297	**0.471**	0.092	0.118	0.256

To provide unbiasedness of estimation, the quadrant r_Q and Kendall correlation coefficients r_K are transformed by taking $\sin\left(\frac{\pi}{2}\cdot\right)$ of their initial values, whereas the Spearman correlation coefficient r_S is transformed by $2\sin\left(\frac{\pi}{6}\cdot\right)$ (Kendall and Stuart 1963).

All the aforementioned estimators were studied, but in what follows, we exhibit results of the best competitors: the most robust minimax variance and bias *MAD* correlation coefficient is defined by (9.10); the *FQ* estimate is computed by formula (9.6) with the M-estimate $\hat{S}$ of scale defined by the score function (9.23); the limit case of the minimax variance M-estimator r_{MMAD} defined at the end of

Sects. 3.2, 3.4 (the most V-robust estimator); the minimum covariance determinant (MCD) estimator (Rousseeuw 1984) of correlation was computed by the means of the package R.

The Monte Carlo estimate mean squared error (MSE) is computed as follows: $MSE(\hat{\rho}) = M^{-1}\sum_{k=1}^{M}(\hat{\rho}-\rho)^2$, where M is a number of trials ($M = 50000$). The relative estimate efficiency (RE) is defined as the ratio of the asymptotic variance of the sample correlation coefficient r and the experimental estimate $\hat{\rho}$ variance: $\mathrm{RE}(\hat{\rho}) = (1-\rho^2)^2/(n\,var(\hat{\rho}))$.

The best performances in table rows are boldfaced, the next to them values are starred.

9.4.2 Discussion

Normal Distribution From Table 9.1 it follows that

1. on small and large samples, the best estimate among the chosen set of robust alternatives to the sample correlation coefficient is the r_{MFQ} M-estimate; the next to it in performance is the sample correlation coefficient r;
2. the Kendall correlation coefficient r_K is the best in performance among the nonparametric measures, especially on small samples;
3. on large samples, estimate biases can be neglected, but not their variances.

Contaminated Normal Distributions From Table 9.2 it follows that

1. the sample correlation coefficient r is catastrophically bad under contamination;
2. on small and large samples, the r_{MCD} correlation coefficient is the best with respect to bias; the next to it is the most B- and V-robust MAD correlation coefficient, and as the computational complexity of r_{MCD} is much higher than that of r_{MAD}, the latter is preferable.
3. the set of performed experiments does not allow to sort out the other estimates, namely r_{FQ} and r_{MFQ}.

The quadrant correlation coefficient r_Q is also an asymptotically minimax bias estimator of the correlation coefficient (Huber 1981) as r_{MAD}. Nevertheless, as it follows from Tables 9.1, 9.2, 9.3 and 9.4, its overall performance is inferior to the performance of that estimator. This can be explained by the choice of the class of direct robust counterparts of the sample correlation coefficient (Huber 1981; Shevlyakov and Vilchevsky 2002b) at which the minimax property of r_Q is established—the class of estimators based on principal variable variances is richer and more advantageous than the competing class. However, we may cautiously recommend the quadrant coefficient r_Q as a moderate robust alternative to the sample correlation coefficient r both due to its low-complexity and to its finite sample binomial distribution (Blomqvist 1950).

Furthermore, the minimax variance and bias r_{TRIM} and r_{MTRIM} estimates computed at $\varepsilon = 0.1$ were inferior in performance not only to their limit cases r_{MAD} and r_{MMAD}-estimates, respectively, but also to r_{FQ} and r_{MFQ} estimates. This can be explained by the fact that the highly robust and efficient FQ_n estimator of scale dominates over Huber's robust minimax variance trimmed standard deviation estimator (Smirnov and Shevlyakov 2014).

Bivariate Cauchy Distributions From Tables 9.3 and 9.4 it follows that

1. the sample correlation coefficient r is again catastrophically bad at both heavy-tailed distributions, especially at the ICD Cauchy density; however, we expected a much worse performance with respect to estimate's mean, as the population means do not exist in this case;
2. on small and large samples, the r_{FQ} correlation coefficient is the best with respect to all performance characteristics; the next to it is the most B- and V-robust MAD correlation coefficient at the ICD Cauchy density and the nonparametric Kendall correlation at the bivariate Cauchy t-distribution;
3. the aforementioned advantages of the MCD correlation coefficient in the contaminated normal case disappear at these heavy-tailed distributions;
4. it seems that the bivariate independent component Cauchy distribution poses more challenges for estimation of correlation as compared to the bivariate Cauchy t-distribution.

9.5 Conclusions

Our main contributions to robust estimation of the correlation coefficient are as follows: (1) a new class of M-estimators of the correlation coefficient for bivariate ICD densities is proposed as the generalization of the maximum likelihood estimator of the correlation coefficient for the bivariate normal distribution generally enhancing the efficiency of estimation as compared to earlier results at slightly contaminated bivariate normal distributions; (2) the comparative performance of various robust estimators of the correlation coefficient is studied at new heavy-tailed bivariate ICD Cauchy densities.

In addition, minimax variance (in Huber's sense) M-estimators of the correlation coefficient for bivariate ε-contaminated normal distributions are designed.

In Monte Carlo experiment, the proposed r_{MFQ} and r_{MMAD} M-estimators proved to be the best on small and large samples at slightly contaminated normal distributions.

Finally, at heavy-tailed bivariate distributions, our earlier proposed r_{FQ} estimator of the correlation coefficient based on the fast highly robust and efficient Rousseeuw–Croux FQ_n estimator of scale can be recommended.

Acknowledgements Here it is the right place to thank Prof. Hannu Oja who recommended one of the coauthors of this paper to use the term "independent component distributions" for the corresponding family of bivariate distributions. Also we thank him for the support of our research on robust correlation.

Furthermore, we thank the reviewers whose comments and remarks helped us much to improve our paper.

References

Blomqvist, N.: On a measure of dependence between two random variables. Ann. Math. Stat. **21**(4), 593–600 (1950)

Devlin, S.J., Gnanadesikan, R., Kettenring, J.R.: Robust estimation and outlier detection with correlation coefficients. Biometrika **62**(3), 531–545 (1975)

Galton, F.: Family likeness in stature. Proc. R. Soc. Lond. **40**, 42–73 (1886)

Gnanadesikan, R., Kettenring, J.R.: Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics **28**(1), 81–124 (1972)

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.: Robust Statistics: The Approach Based on Influence Functions. Wiley, New York (1986)

Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Stat. **35**(1), 73–101 (1964)

Huber, P.J.: Robust Statistics. Wiley, New York (1981)

Kendall, M.G., Stuart, A.: The Advanced Theory of Statistics. Inference and Relationship. Griffin, London (1963)

Pasman, V.R., Shevlyakov, G.L.: Robust methods of estimation of a correlation coefficient. Autom. Remote Control **48**, 332–340 (1987)

Pearson, K.: Contributions to the mathematical theory of evolution. Philos. Trans. R. Soc. Lond. **185**, 77–110 (1894)

Pearson, K.: Notes on the history of correlations. Biometrika **13**, 25–45 (1920)

Pearson, K.: Karl Pearson's Early Statistical Papers. Cambridge University Press, Cambridge (1948)

Rousseeuw, P.J.: Least median of squares regression. J. Am. Stat. Assoc. **79**(388), 871–880 (1984)

Rousseeuw, P.J., Croux, C.: Alternatives to the median absolute deviation. J. Am. Stat. Assoc. **88**(424), 1273–1283 (1993)

Shevlyakov, G.L., Smirnov, P.O.: Robust estimation of the correlation coefficient: an attempt of survey. Austrian J. Stat. **40**(1/2), 147–156 (2011)

Shevlyakov, G.L., Vilchevski, N.O.: Robustness in Data Analysis: Criteria and Methods. VSP, Utrecht (2002a)

Shevlyakov, G.L., Vilchevsky, N.O.: Minimax variance estimation of a correlation coefficient for epsilon-contaminated bivariate normal distributions. Stat. Probab. Lett. **57**, 91–100 (2002b)

Shevlyakov, G.L., Smirnov, P.O., Shin, V.I., Kim, K.: Asymptotically minimax bias estimation of the correlation coefficient for bivariate independent component distributions. J. Multivar. Anal. **111**, 59–65 (2012)

Smirnov, P.O., Shevlyakov, G.L.: Fast highly efficient and robust one-step M-estimators of scale based on FQ_n. Comput. Stat. Data Anal. **78**, 153–158 (2014)

Stigler, S.M.: The History of Statistics. Harvard University Press, Cambridge (1986)

Tukey, J.W.: A survey of sampling from contaminated distributions. Contributions to Probability and Statistics, vol. 2, pp. 448–485. Stanford University Press, Stanford (1960)

Chapter 10
Robust Coordinates for Compositional Data Using Weighted Balances

Peter Filzmoser and Karel Hron

Abstract Multivariate observations which carry exclusively relative information are known under the name *compositional data*, and they have very specific geometrical properties. In order to represent them in the usual Euclidean geometry, they need to be expressed in orthonormal coordinates prior to their possible further statistical processing. As it is not possible to construct Cartesian coordinates for the compositions, that would assign a coordinate for each of the parts separately, a choice of interpretable orthonormal coordinates is of particular interest. Although recent experiences show clear advantages of such coordinates, where the first coordinate aggregates information from logratios with a particular compositional part of interest, their usefulness is limited if there are distortions like rounding errors or other data problems in the involved parts. The aim of the paper is thus to introduce a "robust" version of these coordinates, where the role of the remaining parts (with respect to the part of interest) is weighted according to their relevance for the purpose of the statistical analysis. Theoretical considerations are accompanied by examples with data sets from chemistry and geochemistry, pointing out the role of robust estimation in the context of regression with compositional covariates.

Keywords Compositional data • Isometric logratio coordinates • Variation matrix • MM-regression

P. Filzmoser (✉)
Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, 1040 Vienna, Austria
e-mail: P.Filzmoser@tuwien.ac.at

K. Hron
Department of Mathematical Analysis and Applications of Mathematics, Palacký University, 77146 Olomouc, Czech Republic
e-mail: hronk@seznam.cz

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_10

10.1 Introduction

Compositional data as observations carrying only relative information are represented usually in form of proportions or percentages which allows for a better interpretability. Since the publication of the paper (Pearson 1897) on spurious correlation, a lot of discussion within and also outside the statistical community has evolved about the proper treatment of such data. There have been groups who claimed that no real problem occurs, while others developed complex models to overcome the effect of the unity constraint—in case of a proportional representation of the compositions. However, there were also alternative approaches for the treatment of these particular multivariate observations that frequently occur in many applications, like in natural and social sciences (including biology, geochemistry, metabolomics, econometrics, and many other fields). One such approach was established at the end of the last century, and published in the seminal book on the statistical analysis of compositional data (Aitchison 1986). In this book, possible problems for the analysis of compositional data were pointed out if they were interpreted just as observations with a constant sum constraint (like 1 in proportions and 100 in percentages). The author claimed that all the relevant information in compositional data is contained in the ratios between the variables (so-called compositional parts), or even better, in logratios that are mathematically easier to handle—the logratio methodology was born. From this perspective, proportional or percentage observations are just proper presentations of the generalized concept of compositional data that are from their nature scale invariant. Later on, requirements on a meaningful analysis of compositional data were summarized into mathematical principles (Egozcue 2009) that are formally followed by the Aitchison geometry (Billheimer et al. 2001; Egozcue and Pawlowsky-Glahn 2006; Pawlowsky-Glahn and Egozcue 2001).

In order to express compositions in the real space, on which most of the multivariate statistical methods rely (Eaton 1983), coordinates with respect to the Aitchison geometry need to be constructed. As all information in compositional data is contained in logratios, one can expect that also coordinates capturing the multivariate information of compositional data will be formed of logratios and their linear combinations. Indeed, in this way it is possible to construct different coordinate systems with a specific and desired interpretation, hereby considering the natural requirement of orthonormality (i.e., coordinates with respect to an orthonormal basis) (Egozcue et al. 2003).

Due to the lack of a canonical basis within the Aitchison geometry, it is not possible to assign coordinates to each of the compositional parts simultaneously in one coordinate system. On the other hand, it is possible to construct such orthonormal coordinates that allow to assign one coordinate (or a set of coordinates) to a specific compositional part (or to a group of parts) of interest. Nevertheless, note that these coordinates should not be confounded with specific compositional parts as they describe the relative behavior of a compositional part with respect to all the others. By following the principles of compositional data analysis, where logratios form the elemental source of information in compositions, it is clear that such a

coordinate has to be formed by a sum of logratios of that part with all the remaining parts. Consequently, when some of those parts are strongly affected by measurement errors, this has an impact also to the resulting coordinate itself and may lead to a biased model estimation. Moreover, due to the relative scale of compositional data (i.e., ratios instead of absolute differences are contained in the Aitchison distance), parts with small values, predisposed to be influenced by the accuracy of the measurement device, can be very dangerous in this context. Particularly, small (relative) values of compositional parts are a source of outlyingness of the observations that later on leads to a distortion of classical statistical methods, when they are applied to the coordinate representation of compositions (Filzmoser and Hron 2011).

The aim of this contribution is to build interpretable orthonormal coordinates that downweight the influence of possibly noisy variables and to use them for the further statistical analysis. In the next section, the concept of a weighted balance is introduced and illustrated. Section 10.3 is devoted to the construction of an orthonormal basis of weighted balances. The system of weighted balances serves as explanatory variables in the context of robust regression in Sect. 10.4. The final Sect. 10.5 summarizes and concludes.

10.2 Logcontrasts and Weighted Balances

Compositional data are underlying the Aitchison geometry (Egozcue and Pawlowsky-Glahn 2006; Pawlowsky-Glahn and Egozcue 2001) that fulfills the above principles of compositional data analysis, like scale invariance and relative scale, but also the mathematical concept of subcompositional dominance (see Egozcue 2009, for details). Due to the scale invariance property, any basis of D-part compositional data $\mathbf{x} = (x_1, \ldots, x_D)'$ consists of $D - 1$ elements, what excludes the existence of a canonical basis within the Aitchison geometry. Consequently, the construction of interpretable logratio coordinates is of primary interest. An arbitrary coordinate is a logcontrast (Aitchison 1986), i.e. a term of the form

$$a_1 \ln x_1 + \ldots + a_D \ln x_D = \mathbf{a}' \ln(\mathbf{x}), \qquad \text{where } \sum_{j=1}^{D} a_j = 0. \tag{10.1}$$

We refer to a standard logcontrast if $\mathbf{a}'\mathbf{a} = 1$. Geometrically, vectors $\mathbf{a} = (a_1, \ldots, a_D)'$ are elements of the hyperplane $\mathscr{H} = \{\mathbf{a} \in \mathbf{R}^D : a_1 + \ldots + a_D = 0\}$. Thus, in order to derive orthonormal coordinates, that are usually preferable as they ensure isometry between the Aitchison geometry and the Euclidean real space, it is sufficient to construct $D-1$ orthonormal logcontrasts $\mathbf{a}_i' \ln(\mathbf{x})$, where the requirement of orthonormality is transferred to coefficient vectors $\mathbf{a}_i$, $i = 1, \ldots, D-1$. One such choice as a special case of a more general concept of balances (Egozcue and Pawlowsky-Glahn 2005) leads to vectors

$$\mathbf{a}_i = \sqrt{\frac{D-i}{D-i+1}} \left(0, \ldots, 0, 1, -\frac{1}{D-i}, \ldots, -\frac{1}{D-i}\right)' \tag{10.2}$$

with $i-1$ zero entries that result in $D-1$ orthonormal coordinates

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{k=i+1}^{D} x_k}}, \; i = 1, \ldots, D-1 \tag{10.3}$$

(Egozcue et al. 2003; Hron et al. 2010; Fišerová and Hron 2011). Interestingly, the compositional part x_1 is contained just in the first coordinate, z_1, in form of a (scaled) logratio of that part with the geometric mean of the remaining parts. In other words, this coordinate can be interpreted in terms of relative abundance of x_1 compared to an "average behavior" of the remaining parts in the composition. Alternatively, we can also refer to a scaled sum

$$z_1 = \frac{1}{\sqrt{D(D-1)}} \left(\ln \frac{x_1}{x_2} + \ldots + \ln \frac{x_1}{x_D} \right)$$

of logratios containing the part x_1. In any case, we can conclude that z_1 extracts all relative information concerning x_1. Of course, by permuting the parts in the original composition such that x_l fills the first position, $\mathbf{x}^{(l)} = (x_1^{(l)}, \ldots, x_D^{(l)})' = (x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)$, $l = 1, \ldots, D$, Eq. (10.3) remains formally unchanged, just moving to coordinates $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})'$ with parts from $\mathbf{x}^{(l)}$ instead of $\mathbf{x}$. This permutation guarantees an analogous interpretation of the relation between $z_1^{(l)}$ and x_l as it was previously the case for z_1 and x_1. In other words, in these D coordinate systems always one of the coordinates is assigned to one specific compositional part, which is advantageous for interpretation purposes. Nevertheless, care must be taken, as they cannot be interpreted as representations of just one part, but of the relative contribution of that part with respect to all the others. Together with the fact that different orthonormal coordinate systems are just rotations of each other (Egozcue et al. 2003), the coordinates defined in (10.3) were recently successfully applied to many problems in compositional data analysis (Hron et al. 2010, 2012; Buccianti et al. 2014; Kalivodová et al. 2015).

Nevertheless, a detailed numerical inspection performed for the GEMAS data set, resulting from a large geochemical mapping project covering most of the European countries (Reimann et al. 2012), revealed also problems that may occur when coordinates (10.3) are uncritically accepted for extracting univariate information about single compositional parts. Namely, measurement devices work usually in an absolute scale and just small relative values of components are often burdened by rounding errors or other data imprecisions, which can have a negative effect for the subsequent statistical processing. Consequently, this can result also in a biased view on the role of outlying observations (and variables) in the compositional data set. Moreover, the affected compositional parts can strongly influence the value of the geometric mean in $z_1^{(l)}$, $l = 1, \ldots, D$, and thus also the overall view on the role of single parts in the composition, for which the coordinates are constructed. From

this perspective, it seems to be more meaningful to move to a weighted sum of logratios,

$$\alpha_2 \ln \frac{x_1}{x_2} + \ldots + \alpha_D \ln \frac{x_1}{x_D}, \quad \alpha_2 + \ldots + \alpha_D = 1. \tag{10.4}$$

After rescaling to a standard logcontrast, this leads to the new coordinate

$$z_1^* = \frac{1}{\sqrt{1 + \sum_{k=2}^{D} \alpha_k^2}} \ln \frac{x_1}{\prod_{k=2}^{D} x_k^{\alpha_k}}, \quad \alpha_2 + \ldots + \alpha_D = 1 \tag{10.5}$$

capturing just *relevant* relative information about a compositional part of interest (here, without loss of generality, x_1 was set to be such a part). Using the above notations (10.1) and (10.4), a vector of logcontrast coefficients

$$\mathbf{a} = (a_1, a_2, \ldots, a_D)' = \frac{1}{\sqrt{1 + \sum_{k=2}^{D} \alpha_k^2}} (1, -\alpha_2, \ldots, -\alpha_D)' \tag{10.6}$$

can be assigned to z_1^* (for later purposes, we assign also $a_1 \equiv \alpha_1 = 1$). Note that for $a_2 = \ldots = a_D = -\frac{1}{D}$ (up to a scaling constant) we would obtain exactly the coordinate z_1 from (10.3).

10.2.1 Example 1: Zero-Weights

The GEMAS project mentioned above (Reimann et al. 2012) resulted in a comprehensive geochemical atlas (Reimann et al. 2014a,b), where the geochemistry of agricultural soil in Europe is mapped. The source data for this atlas are available on an attached CD-ROM. For illustrating the above ideas, we use the XRF (X-ray fluorescence) data from the aqua regia (AR) extraction, a relatively strong acid attack where only a certain part of the total element content of the samples will be dissolved. Overall, the concentrations of 41 chemical elements are available, and additionally LOI (loss on ignition) is measured. This data part is a classical case for compositional data because when measuring all the components, their sum is 100 %.

In the following example, the interest is in mapping the element Aluminium (Al). Therefore, a balance is constructed according to Eq. (10.3) where Al is in the role of x_1. The corresponding balance z_1 expresses then all the relative information to the other parts in the composition. Out of the 41 available chemical element concentrations, however, for some variables a high proportion of measurements is below the detection limit. For example, for the elements Bi, Mo, Sb, Sn, Ta, and W,

more than 90 % of the reported data are below the corresponding detection limits. It would not make much sense to consider those variables in the analysis, but a general question is, for which percentage below detection a variable can still be considered for a statistical analysis. For the purpose of mapping Al, we compare the resulting balances if this percentage is zero, and if at most 10 % of the values for the variables may be below detection. In the first case we have 14 variables, in the second case 24 (in addition to Al). Following the notation from above, our "full" composition has $D = 25$ parts, from which the unweighted balance for Al is constructed. The weighted balance has weight $\alpha_1 = 1$ (for Al), and weights α_j (for $j = 2, \ldots, D$) which are either zero (if for the corresponding variable the percentage of values below detection limit is positive) or positive otherwise, considering the condition $\alpha_2 + \ldots + \alpha_D = 1$. In our case, the values for the non-zero weights are 1/14.

Figure 10.1 shows the comparison of the resulting balances: The left map shows the balance for Al using all variables with at most 10 % of values below detection, and the right map puts zero-weights on those variables which contain values below the detection limit. At a first glance, the difference between the two maps seems to be rather small. Figure 10.2 investigates these differences in more detail. The left picture compares both balances in a scatterplot. The dashed line reveals the joint linear trend in the plot, and we are interested in deviations orthogonal to this trend. These orthogonal deviations are visualized in the map on the right-hand side. Here it is immediate that the differences are pronounced mainly in the northern part of Europe. Indeed, this area shows much higher values in the left map of Fig. 10.1 than in the right map (compared to the corresponding values in the other regions). In other words, the variables with detection limit problems led to an increase of the values in the north. It is, however, unclear, whether this increase is based on the problematic values below detection, or on the general data structure of these variables.

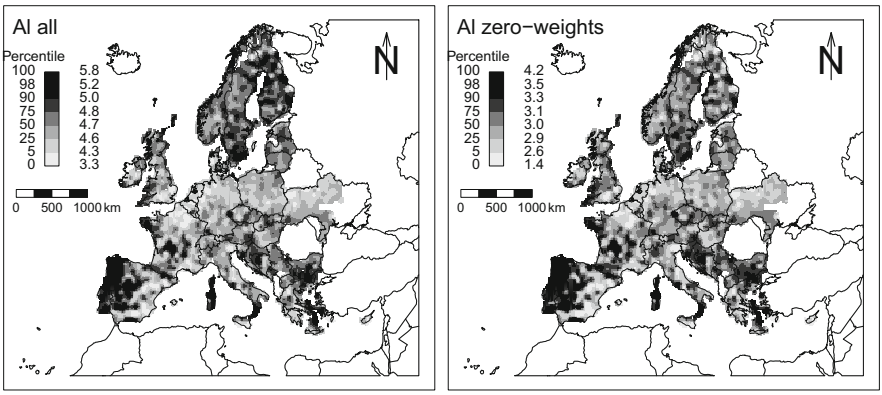

Fig. 10.1 Balances for Aluminium (Al) represented in terms of maps. *Left*: all variables with at most 10 % of values below detection are used to construct the balance; *right*: variables containing any values below detection receive weight zero for the balance

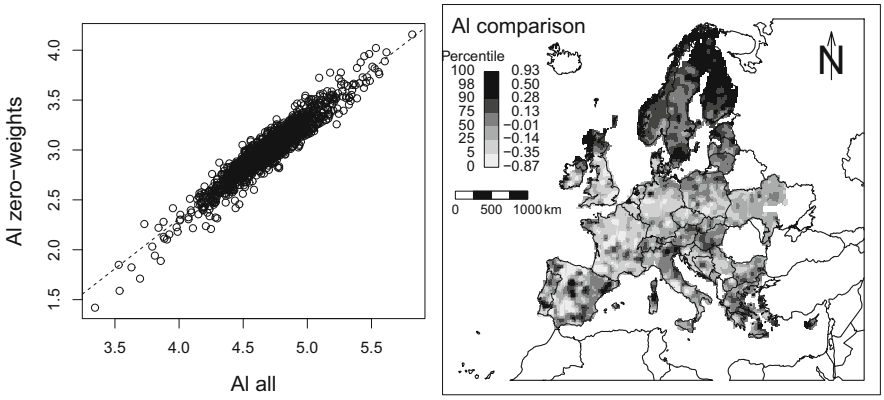

Fig. 10.2 Comparison of the two balances used for mapping Al in Fig. 10.1. *Left*: scatterplot of the two balances, with main trend indicated by the *dashed line*; *right*: orthogonal deviations from the *dashed line* are mapped, indicating the difference between both maps in Fig. 10.1

10.2.2 Example 2: Weights Based on the Variation Matrix

A basic measure of variability of a composition $\mathbf{x} = (x_1, \ldots, x_D)'$ is the variation matrix, defined as

$$\mathbf{T} = \left\{ \mathrm{var}\left(\ln \frac{x_i}{x_j} \right) \right\}_{i,j=1}^{D} \tag{10.7}$$

(Aitchison 1986), where "var" denotes the variance. The elements of the variation matrix describe the variability of the logratio $\ln \frac{x_i}{x_j}$. The smaller the value of this variance, the more the logratio tends to be a constant, and the two parts can be considered as having very similar structure. On the other hand, for large variances the parts are very different to each other—in a non-compositional context one would say that they are uncorrelated. With respect to the choice of appropriate weights $\alpha_2, \ldots, \alpha_D$ for our balance of interest, one could use this information of the variation matrix for defining those weights: large elements in the variation matrix should lead to small weights, and small elements to larger weights. In that way it is possible to downweight variables with a weak relation to our variable of interest.

Here we propose the following choice. Without loss of generality we assume that the part of interest is in the first row (column) $\mathbf{t}_1$ of the variation matrix $\mathbf{T}$ from Eq. (10.7), i.e.

$$\mathbf{t}_1 = (t_{11}, \ldots, t_{1D}) = \left(\mathrm{var}\left(\ln \frac{x_1}{x_1} \right), \mathrm{var}\left(\ln \frac{x_1}{x_2} \right) \ldots, \mathrm{var}\left(\ln \frac{x_1}{x_D} \right) \right).$$

Then

$$\tilde{\alpha}_j = 1 - \frac{t_{1j}}{\max \mathbf{t}_1} \qquad \text{for } j = 1, 2, \ldots, D, \tag{10.8}$$

defines weights in the interval [0, 1]. The final weights satisfying the condition $\alpha_2 + \ldots + \alpha_D = 1$ are then taken as

$$\alpha_j = 1 - \frac{\tilde{\alpha}_j}{\sum_{k=2}^{D} \tilde{\alpha}_k} \qquad \text{for } j = 2, \ldots, D. \tag{10.9}$$

In the interest of robust estimation, we employ a robust estimator for "var," namely the square of the Median Absolute Deviation (MAD).

An application of this choice of weights for the example data set from Sect. 10.2.1 is presented in the following. We consider all compositional parts that do not have any values below detection. Again, the interest is in the part Al, and the balance for Al is compared with the weighted version, using Eq. (10.5), in Fig. 10.3. The left map shows the balance constructed with Al and the remaining 14 variables without values below detection, whereas the right map shows the weighted balance. Figure 10.4 investigates again the differences between both balances—in the same way as in Fig. 10.2. We can find again a clear regional pattern when comparing both balances. Figure 10.4 reveals a clear north–south trend of differences, with the exception of the Scandinavian and Baltic countries. Spain and Portugal also show a very pronounced difference. The weighted balance should be more reliable than its unweighted version because it considers more of the relative information of Al to parts with stronger relation to Al than with weak relation.

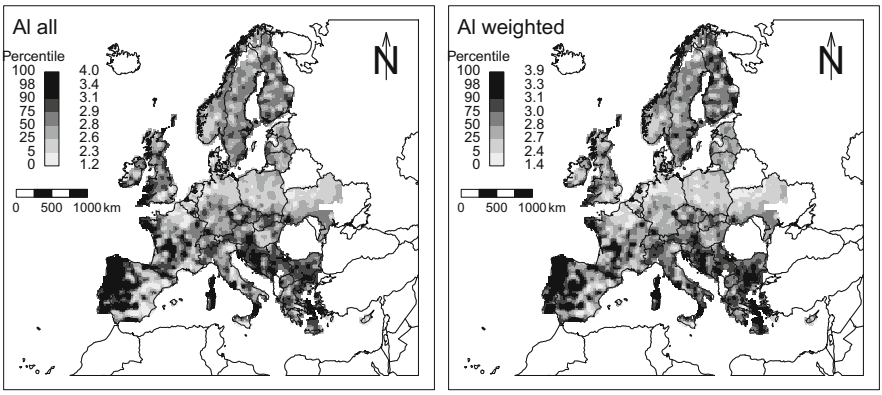

Fig. 10.3 Balances for Aluminium (Al) represented in terms of maps. *Left*: variables without any values below detection are used to construct the balance; *right*: weighted balance for Al with weights based on the variation matrix

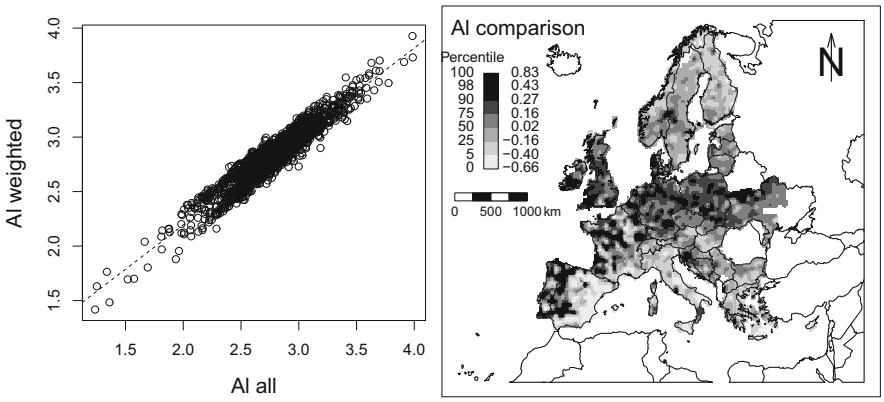

Fig. 10.4 Comparison of the two balances used for mapping Al in Fig. 10.3. *Left*: scatterplot of the two balances, with main trend indicated by the *dashed line*; *right*: orthogonal deviations from the *dashed line* are mapped, indicating the difference between both maps in Fig. 10.3

10.3 Orthogonal Basis of Weighted Balances

In the previous section we were interested in constructing a weighted balance that expresses all the relative information of a particular part. Now we go further and construct a basis with weighted balances.

Let $\alpha_2, \alpha_3, \ldots, \alpha_D$ denote the coefficients of the weighted logratios according to Eq. (10.4), used to construct the first weighted balance. Then, according to Eq. (10.6), the coefficients for the first logcontrast are

$$\mathbf{a}_1 = (a_{11}, \ldots, a_{1D})' = \frac{1}{\sqrt{1 + \sum_{k=2}^{D} \alpha_k^2}} (1, -\alpha_2, \ldots, -\alpha_D)'.$$

For constructing the remaining coordinates $z_2^*, \ldots, z_{D-1}^*$ (or the respective coefficient vectors $\mathbf{a}_i = (a_{i1}, \ldots, a_{iD})'$, $i = 2, \ldots, D-1$) it is sufficient to consider the condition $\mathbf{a}_i' \mathbf{a}_j = \delta_{ij}$, $i, j = 1, \ldots, D-1$ that can be expressed by a sequence of $D-2$ systems of homogenous linear equations,

$$\begin{aligned}
y_1 - \alpha_2 y_2 - \ldots - \alpha_D y_D &= 0, \\
a_{21} y_1 + a_{22} y_2 + \ldots + a_{2D} y_D &= 0, \\
&\vdots \\
a_{i-1,1} y_1 + a_{i-1,2} y_2 + \ldots + a_{i-1,D} y_D &= 0, \\
y_1 + y_2 + \cdots + y_D &= 0,
\end{aligned}$$

for obtaining coefficient vector $\mathbf{a}_i$, $i = 2, \ldots, D-1$. Note that each system contains $D - i$ free parameters what enables to form $D - 3$ coordinates that do not contain the part x_1 by setting the parameter $y_1 = 0$ in the first $D - 3$ equation systems. Together with our previous considerations, z_1^* can be considered to contain the relevant relative information about x_1 and z_{D-1}^* the remaining (redundant) one. Of course, these considerations can be generalized to any of the parts $x_1, \ldots, x_D$ simply by a permutation of parts in the original compositions, in a similar way as it was done for (10.3). As the main aim is to generalize the special choice of balances by assigning weights to single logratios, forming the final logcontrast, we refer to *weighted balances*.

Let us focus in more detail on the special cases $D = 3$ and $D = 4$, where the calculations are still relatively straightforward to show. For $D = 3$, the situation is very easy as there is just one free parameter in the system

$$\begin{aligned} y_1 - \alpha_2 y_2 - \alpha_3 y_3 &= 0, \\ y_1 + y_2 + y_3 &= 0 \end{aligned}$$

to get the orthogonal logcontrast to z_1^* (the orthonormality is achieved afterwards). Consequently, we arrive at coordinates

$$z_1^* = \frac{1}{\sqrt{2(1-\alpha_2\alpha_3)}} \ln \frac{x_1}{x_2^{\alpha_2} x_3^{\alpha_3}}, \quad z_2^* = \frac{1}{\sqrt{6(1-\alpha_2\alpha_3)}} \ln x_1^{\alpha_3-\alpha_2} x_2^{-(1+\alpha_1)} x_3^{1+\alpha_2}. \tag{10.10}$$

This can be compared with the standard (unweighted) balances (Egozcue et al. 2003)

$$z_1 = \sqrt{\frac{2}{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}}, \quad z_2 = \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}. \tag{10.11}$$

For $D = 4$, two homogenous linear systems need to be solved,

$$\begin{aligned} y_1 - \alpha_2 y_2 - \alpha_3 y_3 - \alpha_4 y_4 &= 0, \\ y_1 + y_2 + y_3 + y_4 &= 0 \end{aligned}$$

and consequently also

$$\begin{aligned} y_1 - \alpha_2 y_2 - \alpha_3 y_3 - \alpha_4 y_4 &= 0, \\ (\alpha_4 - \alpha_3) y_2 - (\alpha_4 - \alpha_2) y_3 + (\alpha_3 - \alpha_2) y_4 &= 0, \\ y_1 + y_2 + y_3 + y_4 &= 0, \end{aligned}$$

in order to get the following coordinates (we use the abbreviation $s = \alpha_2\alpha_3 + \alpha_2\alpha_4 + \alpha_3\alpha_4$)

$$
\begin{aligned}
z_1^* &= \frac{1}{\sqrt{2(1-s)}} \ln \frac{x_1}{x_2^{\alpha_2} x_3^{\alpha_3} x_4^{\alpha_4}}, \\
z_2^* &= \frac{1}{\sqrt{2(1-3s)}} \ln x_2^{\alpha_3-\alpha_2} x_3^{\alpha_2-\alpha_4} x_4^{\alpha_4-\alpha_3}, \\
z_3^* &= \frac{1}{\sqrt{16(1-s)(1-3s)}} \\
&\quad \times \ln[x_1^{(\alpha_4-\alpha_3)^2+(\alpha_4-\alpha_2)^2+(\alpha_3-\alpha_2)^2} x_2^{(1+\alpha_3)(\alpha_2-\alpha_3)+(1+\alpha_4)(\alpha_2-\alpha_4)} \\
&\quad \times x_3^{(1+\alpha_2)(\alpha_3-\alpha_2)+(1+\alpha_4)(\alpha_3-\alpha_4)} x_4^{(1+\alpha_2)(\alpha_4-\alpha_2)+(1+\alpha_3)(\alpha_4-\alpha_3)}].
\end{aligned}
$$

Note that in the first system it was possible to set $y_1 = 0$ so that the part x_1 occurs just in the coordinates z_1^* (describing the relevant relative information concerning x_1) and z_3^* (that stands for the redundant one). Similarly, we could proceed also to higher dimensions, nevertheless, growing complexity of the explicit formulas is expected.

As an alternative to the above systems of linear equations, it is possible to consider also the linear relation between the hyperplane $\mathscr{H}$ of vectors of logcontrast coefficients and the real space $\mathbf{R}^{D-1}$ to express them as unconstrained vectors. Consequently, we can use the Gram–Schmidt orthonormalization principle, and a set of orthonormal logcontrasts containing given $\mathbf{a}_1$ is finally obtained by expressing the resulting vectors back in $\mathscr{H}$. On the other hand, by applying this transformation, a direct link of the logcontrast coefficients and the original compositional parts, necessary for the possibility of excluding x_1 from the remaining coordinates (up to one), seems to be lost. Obviously, further research in this direction is needed.

10.4 Robust Regression with Compositional Explanatory Variables

Interpretable orthonormal coordinates are of particular interest for the case of regression with compositional explanatory variables, when the task is to quantify (relative) contributions of single compositional parts to a non-compositional response (Hron et al. 2012). In order to be able to apply statistical inference like hypotheses testing, it is preferable to express the regression model in orthonormal coordinates. Depending on the part of interest for the statistical inference, appropriate orthonormal coordinates need to be constructed. Using the notation from Sect. 10.2, if the interest is in inference about part x_j of the composition $\mathbf{x}$, the corresponding coordinates that

need to be used as explanatory variables are $(z_1^{(j)}, \ldots, z_{D-1}^{(j)})$. In the case $j = 1$, the regression model for a response Y has the following form:

$$\mathsf{E}(Y|\mathbf{x}) = \beta_0 + \beta_1 z_1 + \ldots + \beta_{D-1} z_{D-1}.$$

The parameter β_1 that stands for the contribution of the coordinate z_1 to explaining the response will be of primary importance here, since z_1 contains all the relative information about the part x_1 which is of main interest. Also the absolute term parameter β_0 is usually considered for further analysis. Generally, we consider D regression models

$$\mathsf{E}(Y|\mathbf{x}) = \beta_0^{(j)} + \beta_1^{(j)} z_1^{(j)} + \ldots + \beta_{D-1}^{(j)} z_{D-1}^{(j)}, \quad j = 1, \ldots, D, \tag{10.12}$$

where just the parameters $\beta_0^{(j)}$ and $\beta_1^{(j)}$ are of interest for the interpretation.

Having a concrete realization of the experiment, the regression parameters can be estimated using the standard LS (Least Squares) method. Nevertheless, as both the compositional covariates and the real response can be affected by outlying observations, it is preferable to use robust alternatives. Due to possible changes of the orthonormal coordinate system for the composition $\mathbf{x} = (x_1, \ldots, x_D)'$, just x-affine equivariant estimators are considered. Fortunately, this holds for both LTS (Least Trimmed Squares) (Rousseeuw 1984) and MM estimators (Yohai 1987) that are frequently used in practice. Consequently, for the robust (but also for the classical) estimators $\hat{\beta}_0^{(j)}, \hat{\beta}_1^{(j)}, \ldots, \hat{\beta}_{D-1}^{(j)}$ the following holds:

- $\hat{\beta}_0^{(1)} = \ldots = \hat{\beta}_0^{(D)}$, i.e. the absolute term parameters are equal for any choice of orthonormal coordinates,
- the resulting robust coefficient of determination (R^2) is invariant to orthonormal coordinates of the compositional covariates,
- the same also holds for the F-statistic testing for the significance of the covariates in the regression model.

As mentioned in the previous section, the generalization of balances (10.3) to coordinates $z_1^*, \ldots, z_{D-1}^*$ representing weighted balances can be also extended by a permutation of the parts in the formula (10.5) to $z_1^{*(j)}, \ldots, z_{D-1}^{*(j)}$, $j = 1, \ldots, D$, to stress the role of part x_j in the first coordinate by choosing appropriate weights $\alpha_2^{(j)}, \ldots, \alpha_D^{(j)}$, and used in the regression models (10.12) instead of $z_1^{(j)}, \ldots, z_{D-1}^{(j)}$. By doing so, none of the mentioned properties is lost, the models just suppress the influence of noisy (or similarly affected) parts in the first coordinate that stands for the relevant relative information concerning the compositional part of interest. Additionally, robust regression also allows for a downweighting of outlying observations through the estimation process.

10.4.1 Example 3: Regression on Weighted Balances

We consider a data set that relates to the classification of different types of glass. The response variable is the refractive index, and the oxide content of different elements is available for the use within a regression model as explanatory variables. The complete data set has 214 observations and eight oxides, and it is published at the UCI Machine Learning Database Repository (`http://www.ics.uci.edu/~mlearn/MLRepository.html`). Here we want to illustrate the case of $D = 3$ regressor variables, and thus we use only the oxides Al, Ca, and Si, which form a subcomposition. Moreover, for reasons of illustrating the methodology, all the relative information about Al in the subcomposition is of main interest, and therefore the weighted balances will be constructed with Al in the first position (x_1).

Using the unweighted balances z_1 and z_2 according to the definition (10.11) as explanatory variables in the regression model, we obtain the results shown in Table 10.1 for the classical LS-estimator (snippet of the R summary of the function "lm"). Both balances z_1 and z_2 are significant in the model. z_1 describes all relative information about Al, and z_2 includes the relative information of Ca versus Si.

Now we compare the result with weighted balances, using the definition (10.10). In this example we set the weights $\alpha_2 = 2/3$ and $\alpha_3 = 1/3$, so giving more priority to the relevant information about Al. This choice of the weights is subjective, but it allows to get certain insights into the effect on the results. Table 10.2 shows the outcome. Again, both weighted balances z_1^* and z_2^* are significant in the model. Compared to the results in Table 10.1, the regression coefficient for the first balance has increased (from 0.0018 to 0.0023), which is a consequence of changing the priorities of the logratios $\ln \frac{x_1}{x_2}$ and $\ln \frac{x_1}{x_3}$ in z_1 to weighted versions in z_1^*. The coefficient for the second balance is almost unchanged. Note that the second weighted balance now also includes information about x_1, see (10.10). As mentioned previously, we obtain the same intercept, the same R^2, and the same value for the F-statistic as for the unweighted balances.

The results from classical LS-estimation are compared to robust MM-regression in Table 10.3 (unweighted balances) and Table 10.4 (weighted balances). These tables contain snippets of the summary outputs of the R function "lmrob". We added

Table 10.1 Result of the R function "lm" for classical LS-regression on the unweighted balances

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.5462485  0.0023988 644.584  < 2e-16
z1          0.0017991  0.0003935   4.572 8.22e-06
z2          0.0216312  0.0012246  17.663  < 2e-16

Residual standard error: 0.001654 on 211 degrees of freedom
Multiple R-squared:  0.7063,    Adjusted R-squared:  0.7035
F-statistic: 253.7 on 2 and 211 DF,  p-value: < 2.2e-16
```

Table 10.2 Result of the R function "lm" for classical LS-regression on weighted balances

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.546248   0.002399 644.584  < 2e-16
z1*         0.002321   0.000532   4.363 2.01e-05
z2*         0.021581   0.001171  18.428  < 2e-16

Residual standard error: 0.001654 on 211 degrees of freedom
Multiple R-squared:  0.7063,    Adjusted R-squared:  0.7035
F-statistic: 253.7 on 2 and 211 DF,  p-value: < 2.2e-16
```

Table 10.3 Result of the R function "lmrob" for robust MM-regression on the unweighted balances

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.5382781  0.0025955  592.68   <2e-16
z1          0.0034022  0.0003743    9.09   <2e-16
z2          0.0189489  0.0012559   15.09   <2e-16

Robust residual standard error: 0.0009161
Robust multiple R-squared: 0.8882105
```

Table 10.4 Result of the R function "lmrob" for robust MM-regression on weighted balances

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.5382781  0.0025955 592.681   <2e-16
z1*         0.0002401  0.0005633   0.426     0.67
z2*         0.0192505  0.0011833  16.269   <2e-16

Robust residual standard error: 0.0009161
Robust multiple R-squared: 0.8882105
```

the outcome of a robust R^2 measure defined by Renaud and Victoria-Feser (2010) as

$$R_w^2 = \left(\frac{\sum_{i=1}^n w_i (y_i - \bar{y}_w)(\hat{y}_i - \bar{\hat{y}}_w)}{\sqrt{\sum_{i=1}^n w_i (y_i - \bar{y}_w)^2 \sum_{i=1}^n w_i (\hat{y}_i - \bar{\hat{y}}_w)^2}} \right)^2 \tag{10.13}$$

with $\bar{y}_w = (1/\sum w_i) \sum w_i y_i$ and $\bar{\hat{y}}_w = (1/\sum w_i) \sum w_i \hat{y}_i$, where y_i are the values of the response, and $\hat{y}_i$ are the predictions from the model, for $i = 1, \ldots, n$, and n the number of observations. The weights $w_i = w(r_i, c)$ are directly taken from the fitted model, using a weight function w for the standardized residuals $r_i = (y_i - \hat{y}_i)/\hat{\sigma}$, with the robustly estimated residual standard deviation $\hat{\sigma}$. The weight function used

as default in the function "lmrob" is Tukey's biweight, with the tuning constant $c = 4.685061$ (Maronna et al. 2006).

Also Tables 10.3 and 10.4 show that the estimate for the intercept and the robust R^2 do not change. Note that there is a big difference between the estimated regression coefficient for the first balance in the unweighted and weighted version. This also causes that z_1^* is no longer significant. Of course, we cannot conclude that relative contributions of Al do not play any role for values of the response since Aluminium is now contained also in coordinate z_2^*. Nevertheless, the relevant part of that information (by suppressing the role of Si in favor of Ca) is indeed no more influential.

It may also be interesting to see why the models for LS- and MM-estimation differ. Figure 10.5 shows the comparison of the response with the predicted response, for LS- (left) and MM- (right) estimation. The symbol size in the right plot is related to the inverse weights. Thus, observations with large residuals are downweighted in robust regression. Those observations are also downweighted for the robust R^2. Using the weighted observations in LS-estimation would result in the same coefficient estimates as for MM-regression. Note that the (robust) residuals are the same for the unweighted and the weighted balances.

In order to provide deeper insight into the effect of using different weights for defining the weighted balances, we continuously change the weights α_2 and α_3, considering of course the restriction $\alpha_2 + \alpha_3 = 1$. Figure 10.6 (left) shows on the horizontal axis the ratio α_3/α_2, and the vertical axis represents the corresponding value of the t-statistic for the first weighted balance z_1^*. The results for both LS-regression and MM-regression are shown. Within the two dashed horizontal lines, the test for the corresponding regression parameter does not lead to significance, while outside we obtain significant contributions of this balance to the model.

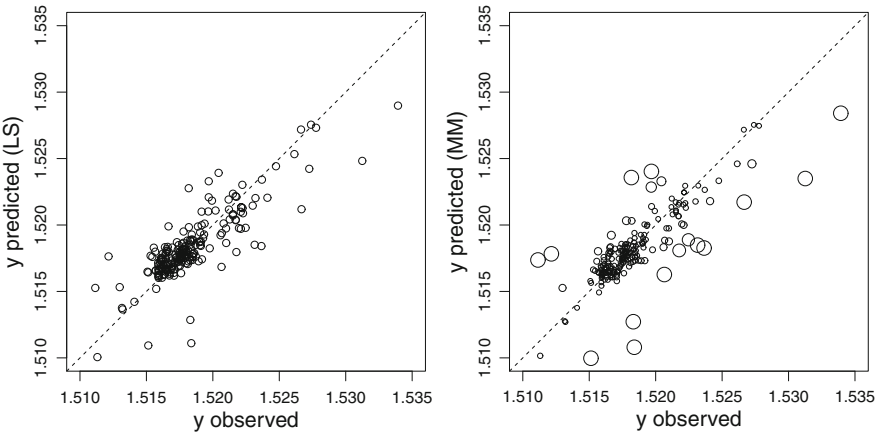

Fig. 10.5 Observed versus predicted response, for LS- (*left*) and MM-estimation (*right*). The symbol size in the *right plot* is inverse proportional to the weights from MM-regression

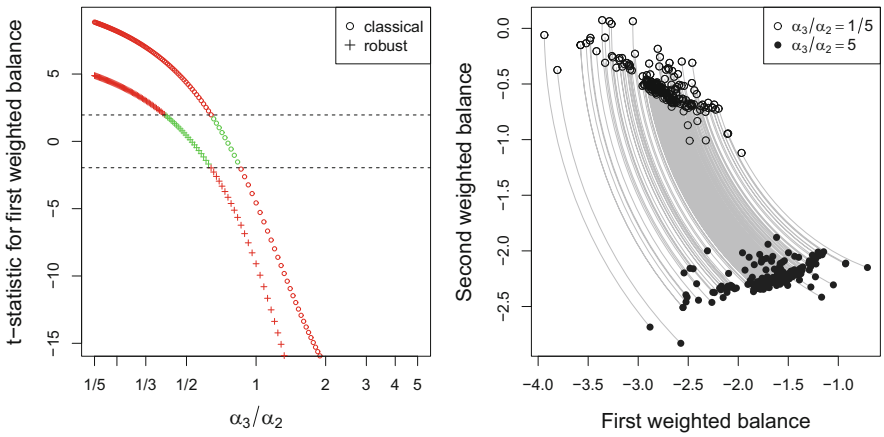

Fig. 10.6 Changing the weights for the weighted balances in terms of a modified ratio α_3/α_2. *Left*: resulting *t*-statistics for z_1^* of classical and robust regression with information about significance (outside the horizontal band); *right*: change in the data structure according to the modified weights

Figure 10.6 (right) shows how the data change in the space of the explanatory variables (z_1^* and z_2^*) according to the weighting scheme. The upper point cloud corresponds to the scheme $\alpha_3/\alpha_2 = 1/5$, and the lower cloud to $\alpha_3/\alpha_2 = 5$. The lines connecting the points are the changes of the data according to the change in the weights—and these correspond to the horizontal axis in the left picture. This figure shows that using different weights for the balances only changes the priorities of the involved parts, but they do not alter the outlyingness of the observations for the regression model. Robustness in regression is thus still important. Even more, the role of outliers can vary among good/bad leverage points and vertical outliers by different weight settings and thus pronounce more or less different results between classical and robust methods as it seems to be also the case in Example 3.

10.5 Summary and Conclusions

Recent experiences with orthonormal coordinates for compositional data, and particularly with balances of type (10.3) that enable to capture all relative information about a part of interest through summing of the corresponding logratios, point out that an automated version of coordinates must not necessary lead to meaningful results. This is due to rounding errors and other effects related to scale invariance and relative scale properties of compositional data that affect the geometric mean in the denominator of the resulting logcontrast. Though, from a theoretical point of view, it is very elegant to have all information about an element "concentrated" in one coordinate, a "robust" alternative to the coordinates (10.3), obtained by assigning weights to the remaining parts in the composition (or the corresponding

logratios), can lead to more reliable results. As the remaining (waste) information about the part of interest can be stored in just one additional coordinate, the resulting orthonormal coordinate system still conveys interpretable outputs. On the other hand, it seems that constructing such coordinates for a general number D of compositional parts calls for quite an effort, as a careful choice of free parameters in systems of linear equations is needed in order to get coordinates that are useful for an interpretation. Moreover, in the regression context also the role of outlying observations (leverage points, vertical outliers) can vary with different choices of elemental weights by preserving the overall multivariate data structure, as different orthonormal coordinate systems are just rotations of each other. We leave concrete answers to these considerations to further research.

Acknowledgements The GEMAS project is a cooperative project of the EuroGeoSurveys Geochemistry Expert Group with a number of outside organizations (e.g., Alterra, The Netherlands; Norwegian Forest and Landscape Institute; Research Group Swiss Soil Monitoring Network, Swiss Research Station Agroscope Reckenholz-Tänikon, several Ministries of the Environment and University Departments of Geosciences, Chemistry and Mathematics in a number of European countries and New Zealand; ARCHE Consulting in Belgium; CSIRO Land and Water in Adelaide, Australia). The analytical work was co-financed by the following industry organizations: Eurometaux, European Borates Association, European Copper Institute, European Precious Metals Federation, International Antimony Association, International Lead Association-Europe, International Manganese Institute, International Molybdenum Association, International Tin Research Institute, International Zinc Association, The Cobalt Development Institute, The Nickel Institute, The (REACH) Selenium and Tellurium Consortium and The (REACH) Vanadium Consortium. The Directors of the European Geological Surveys, and the additional participating organizations, are thanked for making sampling of almost all of Europe in a tight time schedule possible. The Federal Institute for Geosciences and Natural Resourced (BGR), the Geological Survey of Norway and SGS (Canada) are thanked for special analytical input to the project.

The authors gratefully acknowledge the support of the grant COST Action CRoNoS IC1408 and the grant IGA_PrF_2015_013 Mathematical Models of the Internal Grant Agency of the Palacky University in Olomouc.

References

Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman and Hall, London (1986)

Billheimer, D., Guttorp, P., Fagan, W.: Statistical interpretation of species composition. J. Am. Stat. Assoc. **96**, 1205–1214 (2001)

Buccianti, A., Pawlowsky-Glahn, V., Egozcue, J.J.: Variation diagrams to statistically model the behavior of geochemical variables: theory and applications. J. Hydrol. **519**, 988–998 (2014)

Eaton, M.: Multivariate Statistics. A Vector Space Approach. Wiley, New York (1983)

Egozcue, J.J.: Reply to "On the Harker Variation Diagrams;..." by J.A. Cortés. Math. Geosci. **41**, 829–834 (2009)

Egozcue, J.J., Pawlowsky-Glahn, V.: Groups of parts and their balances in compositional data analysis. Math. Geol. **37**, 795–828 (2005)

Egozcue, J.J., Pawlowsky-Glahn, V.: Simplicial geometry for compositional data. In: Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (eds.) Compositional Data in the Geosciences: From Theory to Practice, pp. 145–160. Geological Society, London (2006)

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C.: Isometric logratio transformations for compositional data analysis. Math. Geol. **35**, 279–300 (2003)

Filzmoser, P. Hron, K.: Robust statistical analysis. In: Pawlowsky-Glahn, V., Buccianti, A. (eds.) Compositional Data Analysis: Theory and Applications, pp. 59–72. Wiley, Chichester (2011)

Fišerová, E., Hron, K.: On interpretation of orthonormal coordinates for compositional data. Math. Geosci. **43**, 455–468 (2011)

Hron, K., Filzmoser, P., Thompson, K.: Linear regression with compositional explanatory variables. J. Appl. Stat. **39**, 1115–1128 (2012)

Hron, K., Templ, M., Filzmoser, P.: Imputation of missing valeus for compositional data using classical and robust methods. Comput. Stat. Data Anal. **54**, 3095–3107 (2010)

Kalivodová, A., Hron, K., Filzmoser, P., Najdekr, L., Janečková, H., Adam, T.: PLS-DA for compositional data with application to metabolomics. J. Chemometr. **29**, 21–28 (2015)

Maronna, R., Martin, R.D., Yohai, V.J.: Robust Statistics: Theory and Methods. Wiley, New York (2006)

Pawlowsky-Glahn, V., Egozcue, J.J.: Geometric approach to statistical analysis on the simplex. Stoch. Environ. Res. Risk Assess. **15**, 384–398 (2001)

Pearson, K.: Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc. R. Soc. Lond. **LX**, 489–502 (1897)

Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Dinelli, E., Ladenberger, A.: The concept of compositional data analysis in practice. Total major element concentrations in agricultural and grazing land soils in Europe. Sci. Total Environ. **426**, 196–210 (2012)

Reimann, C., Birke, M., Demetriades, M., Filzmoser, P., O'Connor, P. (eds.): Chemistry of Europe's Agricultural Soils – Part A: Methodology and Interpretation of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B). Schweizerbarth, Hannover (2014)

Reimann, C., Birke, M., Demetriades, M., Filzmoser, P., O'Connor, P. (eds.): Chemistry of Europe's Agricultural Soils – Part B: General Background Information and Further Analysis of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B). Schweizerbarth, Hannover (2014)

Renaud, O., Victoria-Feser, M.-P.: A robust coefficient of determination for regression. J. Stat. Plan. Inference **140**(7), 1852–1862 (2010)

Rousseeuw, P.J.: Least median of squares regression. J. Am. Stat. Assoc. **79**, 871–881 (1984)

Yohai, V.J.: High breakdown-point and high efficiency estimates for regression. Ann. Stat. **15**, 642–65 (1987)

Chapter 11
Computation of the Oja Median by Bounded Search

Karl Mosler and Oleksii Pokotylo

Abstract A new algorithm is given for the exact calculation of the Oja median. It modifies the algorithm of Ronkainen et al. (Developments in Robust Statistics, Springer, Berlin, 2003) by employing bounded regions which contain the median. The regions are built using the centered rank function. The new algorithm is faster and has less complexity than the previous one. It is also used for an even faster approximative calculation.

Keywords Algorithm • Centered rank function • Combinatorial invariance • Multivariate median • Oja outlyingness function

11.1 Introduction

A basic task in multivariate analysis is to describe the general location of data by some point in their middle. Several notions of multivariate medians have been proposed in the literature. They extend different properties and characterizations of the usual univariate median to Euclidean k-space. Besides these defining characterizations the multivariate medians may be distinguished by their invariance properties. These include invariances against monotone transformations of the marginals (like the componentwise median), against spherical transformations (like the spatial median), against affine transformations (like the Oja median, proposed in the seminal paper Oja 1983), and combinatorial invariance. The latter means that the data may be varied in their compartments without changing the median. Examples are the Tukey median (Tukey 1975) and the simplicial median by Liu (1988). These medians are, at least in some sense, more robust against outlying data than the arithmetic mean, which is the center of gravity. Multivariate medians are surveyed by Small (1997) and Oja (2013).

K. Mosler (✉)
Institute of Econometrics and Statistics, University of Cologne, 50923 Köln, Germany
e-mail: mosler@statistik.uni-koeln.de

O. Pokotylo
Cologne Graduate School, University of Cologne, 50923 Köln, Germany
e-mail: pokotylo@wiso.uni-koeln.de

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_11

Like the univariate median most of the multivariate medians can be regarded as maximizers of goal functions, so-called data depths, the Tukey depth, the simplicial depth, the Oja depth, and the spatial depth, among others. See Mosler (2013) for a recent survey.

To be applicable to realistic problems, a median must be computable for dimensions $k > 2$ and at least medium sized data sets. Here we develop an algorithm to calculate the exact value of the Oja median and demonstrate that it is faster, having also less complexity, than the existing ones by Niinimaa et al. (1992) and Ronkainen et al. (2003), ROO hereafter. The exact algorithm can also serve as a benchmark for faster heuristic procedures. In principle, the computation of the Oja median involves repeated checking of all intersections of hyperplanes generated by the data. Our main idea is to introduce bounding hyperplanes that iteratively restrict the area where the median is searched.

The paper is structured as follows: Sect. 11.2 introduces the Oja median and depth and some basic notions and properties connected with them, it also sketches the algorithm of Ronkainen et al. (2003) for exact calculation of the Oja median. In Sect. 11.3 the ideas of the new bounding procedure are discussed, followed by a description of the algorithm in Sect. 11.4. Finally, in Sect. 11.5 numerical experience is reported regarding data in $\mathbb{R}^k$ for k up to dimension seven.

11.2 Oja Median and Depth

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a data set of observations in $\mathbb{R}^k$. Each k observations $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}$ generate an *observation hyperplane* passing through them, which is notated by $p = (i_1, \ldots, i_k)$, $1 \le i_1 < \ldots < i_k \le n$. Let P denote the set of all $\binom{n}{k}$ observation hyperplanes.

k observations together with a given point $\mathbf{x} \in \mathbb{R}^k$ span a simplex in k-space. Its k-dimensional volume is found as

$$V_p(\mathbf{x}) := V(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}, \mathbf{x}) = \frac{1}{k!} abs\left(\begin{vmatrix} 1 & \ldots & 1 & 1 \\ \mathbf{x}_{i_1} & \ldots & \mathbf{x}_{i_k} & \mathbf{x} \end{vmatrix}\right)$$
$$= \frac{1}{k!} abs(d_{0p} + \mathbf{d}_p{}^\top \mathbf{x}).$$

Here d_{0p} is the distance of the hyperplane p from the origin, and $\mathbf{d}_p$ is its normal, given by the vector of cofactors of $\mathbf{x}$ in the determinant. The average of all such volumes is mentioned as the *Oja outlyingness function* of $\mathbf{x}$,

$$\begin{aligned} O(\mathbf{x}|\mathbf{X}) &= ave_{i_1<\ldots<i_k} \left(V(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}, \mathbf{x})\right) \\ &= ave_{i_1<\ldots<i_k} \left(\frac{1}{k!} abs\left(\begin{vmatrix} 1 & \ldots & 1 & 1 \\ \mathbf{x}_{i_1} & \ldots & \mathbf{x}_{i_k} & \mathbf{x} \end{vmatrix}\right)\right) \\ &= \frac{1}{k!} ave_{p\in P} \left(abs(d_{0p} + \mathbf{d}_p^\top \mathbf{x})\right). \end{aligned} \tag{11.1}$$

It is clear from (11.1) that the Oja outlyingness function is piecewise linear and convex on $\mathbf{x}$ as well as continuous on $\mathbf{x}$ and the data in $\mathbf{X}$. The minimizer of the outlyingness function is the *Oja median*, $\mathbf{Med}(\mathbf{X})$. Generally, this median is not unique but forms a convex set. The Oja median is a measure of location and *affine equivariant* regarding $\mathbf{X}$,

$$\mathbf{Med}(\mathbf{Y}) = \mathbf{A}\mathbf{Med}(\mathbf{X}) + \mathbf{b}\,, \tag{11.2}$$

if $\mathbf{Y} = \{\mathbf{A}\mathbf{x}_1 + \mathbf{b}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{b}\}$ with some matrix $\mathbf{A}$ of full rank k and $\mathbf{b} \in \mathbb{R}^k$; see Oja (1983). The outlyingness function can be made affine invariant (to simultaneous transformation of $\mathbf{x}$ and $\mathbf{X}$) by multiplying it with a proper scale factor, *viz.* $(\det \mathbf{S}(\mathbf{X}))^{-1/2}$, where $\mathbf{S}(\mathbf{X})$ is a positive definite $k \times k$ matrix depending on $\mathbf{X}$ and measuring the dispersion of the data cloud $\mathbf{X}$ in an affine equivariant way, that is, with $\mathbf{Y}$ as above, satisfying

$$\mathbf{S}(\mathbf{Y}) = \mathbf{A}^\top \mathbf{S}(\mathbf{X})\mathbf{A}. \tag{11.3}$$

In particular, the usual covariance matrix of $\mathbf{X}$ can serve as $\mathbf{S}(\mathbf{X})$. The *Oja depth function* is defined as (Zuo and Serfling 2000)

$$depth(\mathbf{x}|\mathbf{X}) = \frac{1}{1 + O(\mathbf{x}|\mathbf{X})(\det \mathbf{S}(\mathbf{X}))^{-1/2}}\,. \tag{11.4}$$

Observe that the Oja depth function is affine invariant and continuous. It is maximal at the Oja median of $\mathbf{X}$ and vanishes for $||\mathbf{x}|| \to \infty$. Given $\mathbf{X}$, the depth function is a strictly decreasing transformation of the outlyingness function and, thus, the contour lines of the two functions coincide, though at different values. As the function $O(\cdot|\mathbf{X})$ is convex, all its contour lines are convex. Hence the level sets of the Oja depth are convex and compact sets in $\mathbb{R}^k$. Moreover, the Oja depth decreases monotonically on rays from each point in the median set.

In the case of a centrally symmetric distribution the median set includes the center of symmetry. It can be shown that the Oja depth function determines the data cloud $\mathbf{X}$ uniquely (Koshevoy 2003). The usual breakdown point of the Oja depth is zero, while a slightly different notion of breakdown appears to be positive (Niinimaa et al. 1990).

Given $\mathbf{X}$, the *centered rank function* R is defined by

$$\mathbf{R}(\mathbf{x}) = \frac{1}{k!} ave_{p \in P} \left(S_p(\mathbf{x}) \mathbf{d}_p \right),$$

where

$$S_p(\mathbf{x}) = sign(d_{0p} + \mathbf{d}_p{}^\top \mathbf{x}),$$

indicates on which side of the hyperplane p the point $\mathbf{x}$ is located. Note that $\mathbf{R}(\mathbf{x})$ is the derivative of (11.1), at all $\mathbf{x}$ at which $O(\cdot|\mathbf{X})$ is smooth. Hence, as $O(\cdot|\mathbf{X})$ is convex, the centered rank function is a subgradient of the outlyingness function, at all $\mathbf{x} \in \mathbb{R}^k$. Below, $-\mathbf{R}(\mathbf{x})$ will be used as a direction of descent at point $\mathbf{x}$. It is easily seen from (11.1) that the outlyingness function is also represented as

$$\begin{aligned} O(\mathbf{x}) &= \frac{1}{k!}\left(ave_{p\in P}(S_p(\mathbf{x})d_{0p}) + ave_{p\in P}(S_p(\mathbf{x})\mathbf{d}_p{}^\top\mathbf{x})\right) \\ &= \frac{1}{k!}\frac{1}{\binom{n}{k}}\left(D_0(\mathbf{x}) + \mathbf{D}(\mathbf{x})^\top\mathbf{x})\right), \end{aligned} \tag{11.5}$$

where the sums,

$$D_0(\mathbf{x}) = \sum_{p\in P} S_p(\mathbf{x})d_{0p}\,, \quad \mathbf{D}(\mathbf{x}) = \sum_{p\in P} S_p(\mathbf{x})\mathbf{d}_p\,, \tag{11.6}$$

are piecewise constant. They change by $2d_{0p}$ and $2\mathbf{d}_p$, respectively, when a hyperplane p is crossed.

11.2.1 Calculating the Median According to ROO

In what follows we assume that the data are in general position. Hettmansperger et al. (1999) have shown that a version of the Oja median is always found among the intersection points of observation hyperplanes. The exact algorithm of ROO iteratively optimizes the outlyingness function along the intersection lines of $k-1$ observation hyperplanes, called *observation lines*. At first a searching line is randomly selected among the observation lines and the outlyingness function is optimized along the line. When the point of the minimum is found, the next searching line through this point is chosen. The possible choices of lines depend on the type of the point: the smallest number of lines is obtained if the point is an intersection of hyperplanes that have no common observation points, the largest number is obtained if the point coincides with one of the observation points; see also the discussion before Sect. 11.4.1.

Minimizing the outlyingness function along the searching line is the most time-consuming task. The chosen line L is intersected with all hyperplanes and the outlyingness function (11.1) is calculated at each intersection point. At the first intersection point the constant terms d_{0p} and $\mathbf{d}_p$ are summed up along with the signs $S_p(\mathbf{x}_m)$, yielding the sums D_0, and $\mathbf{D}$ according to (11.6). Then the other intersections are considered step by step. The outlyingness function is calculated as in (11.5). In each new point one of the hyperplanes changes its sign and the sums D_p and $\mathbf{D}$ are updated. Note that there are $\binom{n}{k}$ intersections, almost all of which have to be considered, which causes the great complexity of the algorithm.

11.3 A Bounding Approach

The centered rank function is a subgradient of the outlyingness function. Note that no unique gradient exists at intersections of the observation hyperplanes, hence the centered rank function will in general not vanish at the Oja median. The negative rank function (= negative subgradient) $-\mathbf{R}(\mathbf{x})$ is a vector that points in a direction of descent of the outlyingness function, hence ascent of the depth function. It defines a hyperplane through $\mathbf{x}$, which separates the space into two halfspaces. The positive side of the hyperplane is indicated by the negative subgradient, which equals the negative rank function. Therefore, the Oja median is always found on the positive side of these hyperplanes.

Regarding the Oja depth function, observe that its subgradients have the same direction as the negative subgradients of the Oja outlyingness function,

$$\mathbf{grad}\, \mathit{depth}(\mathbf{x}) = -\mathbf{R}(\mathbf{x})\, (\det \mathbf{S}(\mathbf{X}))^{-1/2}\, (\mathit{depth}(\mathbf{x}))^2\,.$$

Their contour lines coincide since the depth function is a strictly decreasing transform of the outlyingness function.

An example of Oja depth contours and subgradients of the depth function is shown in Fig. 11.1. As expected, all negative subgradients point to the halfspace containing the median, and the gradients are perpendicular to the depth contours.

The halfspaces defined by the negative rank function can be used to build a bounded region that contains the median. In our algorithm we select those halfspaces in an iterative way and restrict the further search to their intersection. The hyperplanes bordering such a search region will be called *bounding hyperplanes*

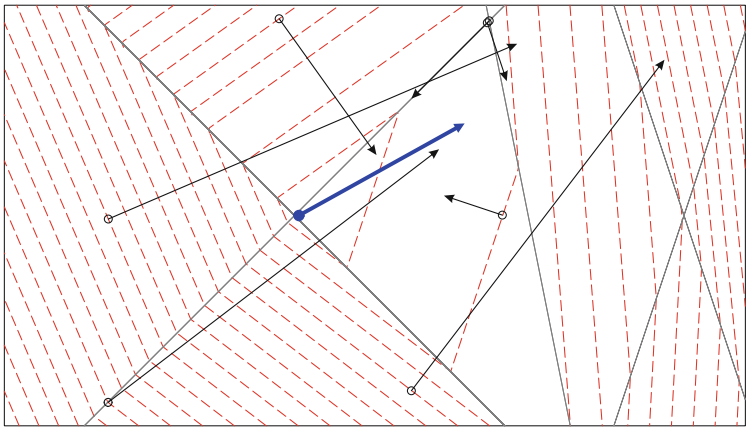

Fig. 11.1 An example of Oja depth contours with values of the negative rank function. The median (unique) is shown at the intersection of the observation lines as a bold point, together with its subgradient

or simply *bounds*. The bounded regions reduce the complexity of the searching procedure by reducing the number of hyperplanes that cross the searching lines as well as the number of their intersections actually considered in the minimization procedure.

The obtained hyperplanes form a bounded region, which is the intersection of the positive sides of the hyperplanes. Actually, such a bounded region is determined by part of these hyperplanes only, as bounds lying outside the region provide no additional information. In our algorithm, we adjust the bounded regions step by step. We begin with a rectangular region limited by hyperplanes that are perpendicular to the coordinate axes and go through the maximal and minimal coordinates of the data points on these axes. Then we add hyperplanes as new bounds. For each added new bound it is checked whether it is *efficient*, that is, actually crosses the bounded region, and thus reduces it. Then the intersection of the new hyperplane with the bounded region is determined, and all bounds that are made inefficient by the new one are removed. To check whether a hyperplane crosses the bounded region, it suffices to check if there exist any two bounds' intersections lying on different sides of the hyperplane. As the calculation of the Oja rank function is itself a rather expensive operation, we will try to obtain the smallest possible central region by performing as few calculation as possible.

We have developed several approaches of the iterative bounds search. The divisive approach (A) is the simplest solution.

Approach A:
The bounded region is iteratively reduced by a divisive approach (A) *viz.* by iteratively adding hyperplanes that go through a properly chosen central point of the region and have their normal vectors equal to the corresponding negative rank function. The central point should be selected to cut a large amount of volume from the bounded region, and shall ideally be the center of the volume, so that any hyperplane through this point will approximately cut off half of the bounded region's volume. Here, we select the mean value of the bounds' intersection points as a central point. As the region is reduced by a hyperplane through the central point, it is expected that its volume shall become (on an average) twice smaller at each step. Ideally, after nine such steps, in any dimension k, a subspace volume of approximately 0.1 % of the initial one should be obtained. The experiments in Sect. 11.5 show that the volumes decrease slower in concrete calculations.

The divisive approach (A) considers only the directions of the subgradients, although their lengths also give the information about the location of the median. Another solution (approach B) consists in moving along the subgradients as it is shown in Fig. 11.2a. The length of $\mathbf{R}(\mathbf{x})$ decreases as $\mathbf{x}$ moves towards the median.

Approach B:

(1) Start with $i = 0$. Select an initial point $\mathbf{x}_0$. Specifically, we choose the componentwise median of all observations.
(2) Determine the subgradient $-\mathbf{R}(\mathbf{x}_i)$.
(3) Add the subgradient vector, $\mathbf{x}_{i+1} = \mathbf{x}_i - \mathbf{R}(\mathbf{x}_i)$, and continue.

We continue building such gradients, each time getting closer to the median, until they become either zero or increase in length. The zero case means that the point $\mathbf{x}_i$ lies in the median set, where the Oja depth assumes its minimal value. As it is seen in Fig. 11.1, the subgradient's length depends not only on the distance from the median, but also on the

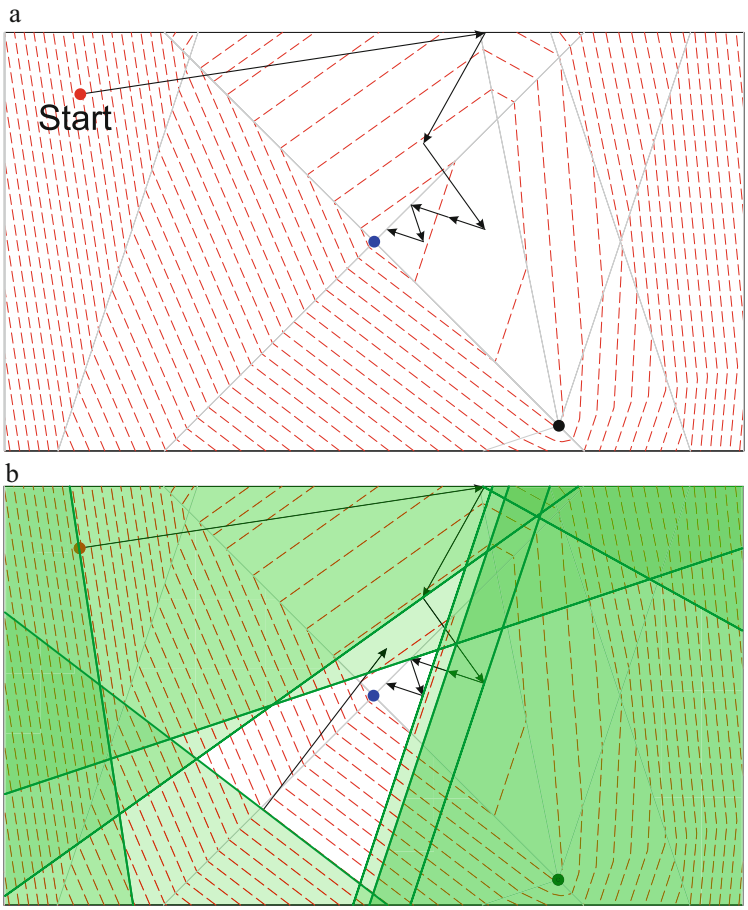

Fig. 11.2 A gradient path (**a**) and a bounded region (**b**), built using the gradients. The subspaces, cut off by each of the bounds are shaded

subspace, formed by hyperplanes, that contains $\mathbf{x}_i$. Thus if the gradients become longer, their lengths may be restricted to the length of the shortest one, and this bound will consequently decrease.

Several of the gradients found may be used to build the bounded searching region, containing the median. The points having shortest gradients are closest to the median. An example of a bounded region built on such gradients is shown in Fig. 11.2b.

The divisive approach needs an almost constant number of calculations to reach the intended volume. However, the efficiency of moving along the subgradients (approach A) strongly depends on the form of the data. In most cases, the subsequent gradients extend in rather different directions, and the volume of the bounded region decreases fast. But in certain cases, especially with asymmetric datasets, this is not true. The subgradients in the sequence may approach the median in a more common

direction and thus leave too much space inside the bounded region. The gradients may also end outside the bounded region or jump between two subsets formed by the observation hyperplanes, providing not much information on each step.

It is therefore reasonable to start with moving along the subgradients, and then, as soon as this procedure slows down, shift to the divisive procedure, until the needed volume is reached:

Approach C:
This yields the following hybrid approach, where the next cutting point may be defined as the end of the subgradient, $\mathbf{x}_{i+1} = \mathbf{x}_i - \mathbf{R}(\mathbf{x}_i)$, as long as it lies inside of the bounded region, or as the center of the bounded region otherwise.

Approach D:
Also the direction of the subgradients can be used to define the next cutting point as a central point of the segment between the subgradient's origin and its intersection with the bound.

Further, the calculation can be accelerated by using rougher bounds, *viz.* enlarging the given bounded region by a circumscribed k-variate box. Then *a fortiori* a point lies outside the bounded region if it lies outside the circumscribed box.

Once a bounded region is defined, the observation hyperplanes lying outside of it are excluded from the searching process, which decreases the number of intersections when minimizing on a line. A problem may occur if the bounded region contains no path through the intersections of the observation lines from the initial searching line to the line containing the median. Such a path connecting any two observation lines may be provided by including the bounds themselves into the searching process as ordinary observation hyperplanes. Figure 11.3 shows an

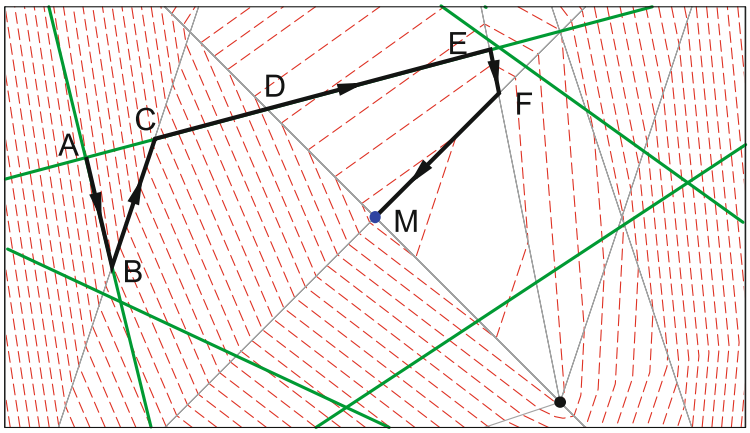

Fig. 11.3 A path through the observation lines (*thin gray*) and the bounds (*bold green*). We start from taking one of the bounds AB as the initial line, and find a minimum point B. Then the outlyingness function is minimized along the next line through this point. As it is seen from the line CE, the point of minimum E is not necessarily the closest one (D) to the median, and the selected path may be not the shortest one. The paths BC and EFM are isolated, as there are no observation lines inside the bounded region to connect them, but they are connected with the bound CE

example of a path from the initial line through the observation lines and bounds to the median.

The bounding method may also be used to find the median in an approximative way with some given precision. The space may be cut until the bounded region has the proper size and its center may be taken as an approximation of the median. It is clear that the median cannot lie outside the bounded region, so its center can be assumed to be the median with precision equal to half of the region's size. As the method considers all observation hyperplanes, it cannot be more efficient than existing approximative methods that consider subsamples of the data.

11.4 The Algorithm

To start with, the first bounded region is created as described in the previous section. The desired size of the bounded region is selected as a part of the original volume. Here the volume is calculated as the volume of a minimal multivariate circumscribed rectangle with edges parallel to the coordinate axes. In subsequent iterations the first bounded region is reduced until the desired volume is reached. Here, the divisive approach (A) is considered, as it shows the best results in experiments (see Sect. 11.5).

Next the initial line is determined. In a two-dimensional space any of the observation lines crossing the bounded region may be selected. In higher dimensions the search of the initial line is more complicated. All intersections of $(k-1)$ hyperplanes are inspected until a first intersection line that crosses the bounded region is found. For this, we start with the lines that border the initial bounded region, which makes the search for a fitting line much easier.

It is clear that all points inside the bounded region lie on the same side of any hyperplane which does not cross this region. Therefore, the respective parts of the sums in (11.6) can be calculated beforehand, which significantly decreases the number of calculations on each step. Thus, on every searching line we may restrict ourselves to iterating the remaining hyperplanes.

The bounded region reduces the procedure of minimization along a line to its part lying inside the region. The searching line is usually intersected by most of the bounds. Therefore the two bounds that cut the bounding region at the intersection line are of primary interest. In order to find these bounds, all bounds are sorted according to their intersections with the searching line. Then the intersection point of the first bound with the searching line is taken as a reference point. The first bound which has the reference point on its positive side is selected as well as the previous one. If the searching line goes through the bounded region, all other bounds must have the reference point on the positive side. This property is used to determine whether a searching line hits the bounded region in dimensions higher than two, as there exist hyperplanes that are crossing the bounded region, but whose intersection line lies outside of it, as it is shown in Fig. 11.4 for a two-dimensional example and in Fig. 11.5 for a higher dimensional one.

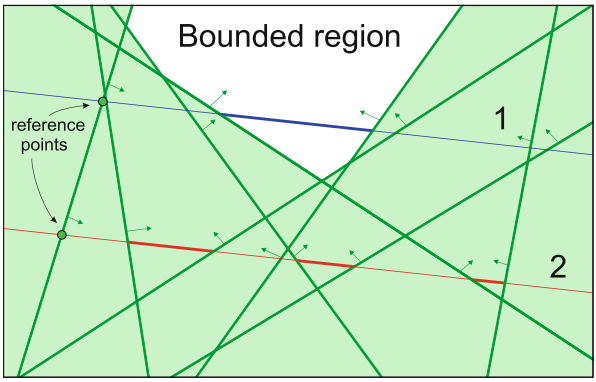

Fig. 11.4 *Two lines*: crossing the bounded region (*1*) and lying outside of it (*2*). The *arrows* show the positive sides of the hyperplanes. The segments between the bounds, one having the reference point on the negative and another one on the positive side, are shown in *bold*. A line that hits the bounded region has only one such segment

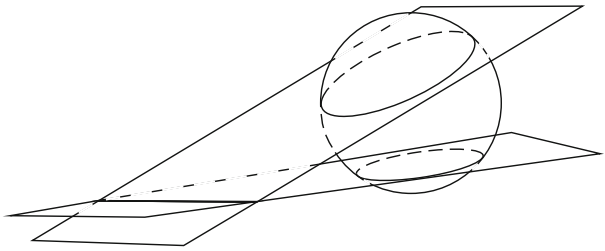

Fig. 11.5 An example of an observation line lying outside of the bounded region (shown as a *sphere*) formed by two hyperplanes crossing the bounded region in a higher dimensional space

The searching line is intersected with the included hyperplanes, and the outlyingness function (11.1) is calculated at every intersection point that lies between the two bounds, found on the previous step. At first, the hyperplanes that intersect the line outside the rougher, i.e. more liberal, bound are filtered out and added to (11.6). Then the first bound's intersection is taken as a median candidate, and as a starting point for the minimization procedure. The left hyperplanes are added to (11.6) with the sign they have in the first bound's intersection. The intersection points are iterated, the corresponding hyperplanes change the sign in the sum (11.6) and the outlyingness function is calculated as in (11.5). The outlyingness function is also calculated at the intersections with the bounds. When the second bound's intersection is reached, the procedure is terminated. The outlyingness function has convex contours and therefore is unimodal on any line. However, in practice the outlyingness function may slightly fluctuate when it is optimized along a line. In this case, as soon as the outlyingness value begins to increase by a certain threshold amount, the minimization along the line is terminated.

When the minimum is found on the searching line, the next observation line is chosen among the lines that contain the minimum. In the simplest case, k hyperplanes, each defined by k unique observation points, define a point at their intersection and produce k observation lines through this point. More complex cases occur when some of the hyperplanes have observation points in common, and their intersection point lies in an affine subspace of dimension $d < k$, generated by these common points. Such a point may then be described by all possible observation hyperplanes that have the same common points, and thus the number of observation lines increases. If the number of observation lines exceeds the predefined maximum number max_{n_L}, ROO propose either to stop, or to take a random subset of these lines. Fischer et al. (2010) in their R-package *OjaNP* used to choose a new initial line in such cases.

If the minimum is defined with one or more bounds, we treat them like ordinary hyperplanes. In order to explicitly determine the bounded region's boarding lines and corners, the bounds are identified by k unique points that are found as intersections with the coordinate axes. If a bound is parallel to some of the axes, the diagonal axes in the space are taken. Thus the bounds do not have identifying points in common, and each intersection of the bounds and observation hyperplanes produces a minimum possible number of observation lines.

11.4.1 Formal Description of the Algorithm

The formal description modifies the one of Ronkainen et al. (2003, A.1, A.2) and includes parts of it to make the comparison easier. In particular, Procedure 1 extends A.1 with the bounded region search (steps 2–14), and Procedure 3 modifies the minimization algorithm A.2 to be used in a bounded region (added steps 1–6, modified steps 18–31). Procedure 2 describes the bounded region construction as in the divisive approach A.

Procedure 1 Compute the exact Oja median.

Input: Data set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in $\mathbb{R}^k$.
The desired size s of the bounded box, $s = \frac{bounded\ box\ volume}{original\ volume}$.
Max number of observation lines to scan max_{n_L}.

Output: Exact Oja median $\mathbf{T} = \mathbf{Med}(\mathbf{X})$.

1: Precalculate all observation hyperplanes $p = (i_1, \ldots, i_k)$, $1 \leq i_1 < \ldots < i_k \leq n$.
2: Build the bounded region $\mathbf{B}$, that is, the set of bounds defining it, using procedure 2.
Chose the initial line L:
3: **for all** subsets $\mathbf{B}_s \subset \mathbf{B}$ with $|\mathbf{B}_s| = k - 1$ **do** ▷ find lines
4: Set $L \leftarrow \bigcap \mathbf{B}_s$.
5: Sort the bounds $\mathbf{b} \in \mathbf{B}$ according to their intersection points with L as ROO do in A.2, i.e. if $L = \{\mathbf{L}_0 + \beta \mathbf{u}_L : \beta \in \mathbb{R}\}$ and we have $\mathbf{b}_i \cap L = \{\mathbf{L}_0 + \beta_i \mathbf{u}_L\}$

and $\mathbf{b}_j \cap L = \{\mathbf{L}_0 + \beta_j \mathbf{u}_L\}$ for some $\mathbf{b}_i, \mathbf{b}_j \in \mathbf{B}$, then $i < j \iff \beta_i < \beta_j$. Denote the order $\mathbf{b}_{(1)}, \mathbf{b}_{(2)}, \ldots, \mathbf{b}_{(nb)}$, where $nb = |\mathbf{B}|$.

Set $\mathbf{y}_1 \leftarrow L \cap \mathbf{b}_{(1)}$.
$i \leftarrow$ smallest i at which $S_{\mathbf{b}_{(i)}}(\mathbf{y}_1) = 1$.
if $\exists j : j > i, S_{\mathbf{b}_{(j)}}(\mathbf{y}_1) \neq 1$ **then**
    Continue ▷ the line is out of bounds
else
    Break ▷ the line is found
end if
end for

Precalculate $\frac{1}{k!}\times$ the common part of (11.6), for given $\mathbf{t}$ in the bounded region:
$\mathbf{H} \leftarrow \sum_{p \notin \mathbf{B}} \frac{1}{k!} S_p(\mathbf{t}) \mathbf{d}_p$,
$H_0 \leftarrow \sum_{p \notin \mathbf{B}} \frac{1}{k!} S_p(\mathbf{t}) d_{0p}$.
Compute $\hat{\mathbf{T}} \leftarrow \arg\min_{\mathbf{t} \in L} O(\mathbf{t})$ using procedure 3.
Set the median candidate $\mathbf{T} \leftarrow \hat{\mathbf{T}}$.
Initialize the collection of investigated lines $\mathscr{L} \leftarrow \{L\}$.
Let n_L be the number of the observation lines containing $\hat{\mathbf{T}}$.
if $n_L > max_{n_L}$ **then**
    There are too many possibilities. Goto 3.
end if
Construct the observation lines $\mathscr{L}' \leftarrow L_1, \ldots L_{n_L}$.
Set $\mathscr{L}' \leftarrow \mathscr{L}' \setminus \mathscr{L}$.
while $\mathscr{L}' \neq \emptyset$ **do**
    Find the line $L \in \mathscr{L}'$ of deepest descent.
    Compute $\hat{\mathbf{T}} \leftarrow \arg\min_{\mathbf{t} \in L} O(\mathbf{t})$ using procedure 3.
    Update $\mathscr{L} \leftarrow \mathscr{L} \cup \{L\}$ and $\mathscr{L}' \leftarrow \mathscr{L}' \setminus \{L\}$
    if $O(\hat{\mathbf{T}}) < O(\mathbf{T})$ **then**
        $\mathbf{T} \leftarrow \hat{\mathbf{T}}$
        Goto 16.
    end if
end while
return T

Procedure 2 Build the bounded region as in the divisive approach A, Sect. 11.3.

Input: Data set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in $\mathbb{R}^k$.
Precalculated observation hyperplanes P.
The desired size s of the bounded box, $s = \frac{bounded\ box\ volume}{original\ volume}$.
Output: The bounded region $\mathbf{B}$.
Enclosing box E.

Define $\mathbf{B} \leftarrow \emptyset$. ▷ the set of bounds
Define $\mathbf{C} \leftarrow \emptyset$. ▷ the set of bounds' intersections
Build the Initial Box:
for $d = 1, \ldots, k$ **do** ▷ index $(-d)$ means all coordinates from 1 to k except of d
    Set the seed of a bound $\mathbf{o}_d \leftarrow \max\{x_{1d}, \ldots, x_{nd}\}$, $\mathbf{o}_{-d} \leftarrow 0$.

Set the normal vector $\mathbf{n}_d \leftarrow -1$, $\mathbf{n}_{-d} \leftarrow 0$.
Define bound $\mathbf{b}$ with $\mathbf{o}$ and $\mathbf{n}$.
ADDBOUND($\mathbf{b}$)
Set the seed of a bound $\mathbf{o}_d \leftarrow \min\{x_{1d}, \ldots, x_{nd}\}$, $\mathbf{o}_{-d} \leftarrow 0$.
Set the normal vector $\mathbf{n}_d \leftarrow +1$, $\mathbf{n}_{-d} \leftarrow 0$.
Define bound $\mathbf{b}$ with $\mathbf{o}$ and $\mathbf{n}$.
ADDBOUND($\mathbf{b}$)
end for ▷ the Initial Box is now built
Proceed with the following divisions:
Calculate the original volume of the space as
$OriginalVolume = \prod_{d=1}^{k} (\max\{x_{1d}, \ldots, x_{nd}\} - \min\{x_{1d}, \ldots, x_{nd}\})$
while $NewVolume/OriginalVolume > s$ **do**
Define the center of $\mathbf{B}$ as $\bar{\mathbf{C}}$.
Calculate the negative rank function $\mathbf{g} = -\mathbf{R}(\bar{\mathbf{C}})$.
Define bound $\mathbf{b}$ with $\bar{\mathbf{C}}$ and $\mathbf{g}$.
ADDBOUND($\mathbf{b}$)
Calculate $NewVolume = \prod_{d=1}^{k} \left(\max\{\mathbf{C}_{1d}, \ldots, \mathbf{C}_{|\mathbf{C}|d}\} - \min\{\mathbf{C}_{1d}, \ldots, \mathbf{C}_{|\mathbf{C}|d}\}\right)$ using the updated intersection points.
end while

function ADDBOUND(new bound $\mathbf{b}$)
▷ here $(\mathbf{b} \cdot \mathbf{x})$ is the dot product of a point $\mathbf{x}$ and the normal vector of $\mathbf{b}$
if (Initial Box is built) and
$(sign(\mathbf{b} \cdot \mathbf{c}_1) = sign(\mathbf{b} \cdot \mathbf{c}_2) \forall \mathbf{c}_1, \mathbf{c}_2 \in \mathbf{C})$ **then**
exit without changes ▷ $\mathbf{b}$ lies outside of $\mathbf{B}$
end if
for all subsets $\mathbf{B}_s \subset \mathbf{B}$ with $|\mathbf{B}_s| = k - 1$ **do** ▷ find new intersections
Set $\mathbf{c} \leftarrow \bigcap(\mathbf{B}_s \cup \mathbf{b})$
if $\forall \mathbf{b} \in \mathbf{B}\ sign(\mathbf{b} \cdot \mathbf{c})! = -1$ **then**
Add the new crossing point $\mathbf{C} \leftarrow \mathbf{C} \cup \mathbf{c}$.
end if
end for
Add the new bound $\mathbf{B} \leftarrow \mathbf{B} \cup \mathbf{b}$.
$\mathbf{C} \leftarrow \mathbf{C} \setminus \{\mathbf{c} : \mathbf{c} \in \mathbf{C}, sign(\mathbf{b} \cdot \mathbf{c}) = -1\}$. ▷ Remove the cut off intersections
$\mathbf{B} \leftarrow \mathbf{B} \setminus \{\mathbf{b} \in \mathbf{B} : sign(\mathbf{b} \cdot \mathbf{c}_1) = sign(\mathbf{b} \cdot \mathbf{c}_2) \forall \mathbf{c}_1, \mathbf{c}_2 \in \mathbf{C}\}$.
end function

Procedure 3 Minimize the outlyingness function O on the chosen line.

Input: Precalculated observation hyperplanes P.
Searching line L.
The bounded region $\mathbf{B}$.
Enclosing box E.

Output: The minimum $\hat{\mathbf{T}} \leftarrow \arg\min_{\mathbf{t} \in L} O(\mathbf{t})$ or an empty point if $L \cap \mathbf{B} = \emptyset$.

Sort bounds $\mathbf{b} \in \mathbf{B}$ according to their intersection points with L as in procedure 1.5, $\mathbf{b}_{(1)}, \mathbf{b}_{(2)}, \ldots, \mathbf{b}_{(nb)}$, where $nb = |\mathbf{B}|$.
Set $\mathbf{y}_1 \leftarrow L \cap \mathbf{b}_{(1)}$.

Set $\mathbf{y}_{b1} \leftarrow L \cap \mathbf{b}_{(i-1)}$ and $\mathbf{y}_{b2} \leftarrow L \cap \mathbf{b}_{(i)}$ where $i = \arg\min_i(S_{\mathbf{b}_{(i)}}(\mathbf{y}_1) = 1)$
if $\exists j : j > i, S_{\mathbf{b}_{(j)}}(\mathbf{y}_1) \neq 1$ **then**
    return empty point. ▷ the line is out of bounds
end if
Choose any point $\mathbf{t}_0 \in \mathbf{B} \cap L$ (e.g. $\mathbf{t}_0 = \mathbf{y}_{b1}$).
Initialize $\mathbf{D} \leftarrow \mathbf{H}$, $D_0 \leftarrow H_0$, $\mathscr{H} \leftarrow \emptyset$.
for all $p \in \mathbf{B}$ **do** ▷ Compute the sum for hyperplanes, crossing L outside of E.
    if $p \cap L \subset E$ **then**
        $\mathscr{H} \leftarrow \mathscr{H} \cup p$
    else
        $\mathbf{D} \leftarrow \mathbf{D} + \frac{1}{k!} S_p(\mathbf{t}_0)\mathbf{d}_p$
        $D_0 \leftarrow D_0 + \frac{1}{k!} S_p(\mathbf{t}_0) d_{0p}$.
    end if
end for
Sort hyperplane indexes $p \in \mathscr{H}$ according to their intersection points with L as ROO do in A.2, $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(np)}$, where $np = |\mathscr{H}|$ and $<$ resp. $\leq$ denote the order of intersection points.
Define $\mathscr{H}_1 \leftarrow \{p : p \in \mathscr{H}, p \cap L < \mathbf{y}_{b1}\}$,
$\mathscr{H}_2 \leftarrow \mathscr{H} \setminus \mathscr{H}_1$,
$\mathscr{H}_3 \leftarrow \{p : p \in \mathscr{H}_2, p \cap L \leq \mathbf{y}_{b2}\}$.
Set $\mathbf{y}_1 \leftarrow L \cap p_{(1)}$ and $\mathbf{y}_{np} \leftarrow L \cap p_{(np)}$.
Compute $\mathbf{D} \leftarrow \mathbf{D} + \sum_{p \in \mathscr{H}_1} \frac{1}{k!} S_p(\mathbf{y}_{np})\mathbf{d}_p + \sum_{p \in \mathscr{H}_2} \frac{1}{k!} S_p(\mathbf{y}_1)\mathbf{d}_p$ and
$D_0 \leftarrow D_0 + \sum_{p \in \mathscr{H}_1} \frac{1}{k!} S_p(\mathbf{y}_{np}) d_{0p} + \sum_{p \in \mathscr{H}_2} \frac{1}{k!} S_p(\mathbf{y}_1) d_{0p}$.
Set potential minimum $\hat{\mathbf{T}} \leftarrow \mathbf{y}_{b1}$.
Evaluate $O(\hat{\mathbf{T}}) = \mathbf{D}^\top \hat{\mathbf{T}} + D_0$.
for all $\{i : p_{(i)} \in \mathscr{H}_3\}$ **do**
    Set $\mathbf{D} \leftarrow \mathbf{D} - \frac{1}{k!} S_{p_{(i-1)}}(\mathbf{y}_1)\mathbf{d}_{p_{(i-1)}} + \frac{1}{k!} S_{p_{(i-1)}}(\mathbf{y}_{np})\mathbf{d}_{p_{(i-1)}}$,
    Set $D_0 \leftarrow D_0 - \frac{1}{k!} S_{p_{(i-1)}}(\mathbf{y}_1) d_{0p_{(i-1)}} + \frac{1}{k!} S_{p_{(i-1)}}(\mathbf{y}_{np}) d_{0p_{(i-1)}}$.
    Set $\mathbf{t} \leftarrow L \cap p_{(i)}$.
    Evaluate $O(\mathbf{t}) = \mathbf{D}^\top \mathbf{t} + D_0$
    if $O(\mathbf{t}) < O(\hat{\mathbf{T}})$ **then**
        Set $\hat{\mathbf{T}} \leftarrow \mathbf{t}$ and $O(\hat{\mathbf{T}}) \leftarrow O(\mathbf{t})$.
    end if
end for
return $\hat{\mathbf{T}}$.

11.5 Numerical Experience and Conclusions

The new algorithm was implemented along the lines of the R-package *OjaNP* of Fischer et al. (2010). A function `ojaMedianExB` was implemented to be used in place of the previous `ojaMedianEx` by ROO. A parameter `alg="exact_bounded"` was added to the function `ojaMedian`, and the corresponding C++ routines were

modified. The codes may be found on the second author's page www.cgs.uni-koeln.de/pokotylo.html. The benchmark values were measured inside the C++ routines, using the file logging. This allows to easily compare the efficiency of the original and modified algorithms, excluding the data transformation and hyperplanes generation time.

The desired volume of the bounded region was set, and the calculation time was determined for the new exact algorithm as well as for the ROO procedure. Best results were received at around 10^{-8} of the original volume in most of tried datasets. The new algorithm showed to be three to six times faster than the one by ROO.

The new algorithm is able to calculate data sets of the same size and dimension as the ROO algorithm. It is mainly restricted by the amount of RAM, as it needs to store all $\binom{n}{k}$ hyperplanes. For example, the calculation of the median in a data set of size 5×100 needs 12 GB RAM. A PC with a 3.4 GHz processor and 32 GB RAM was employed in the experimental studies. Only one processor core was used. The algorithm was able to find the median in data sets of sizes 3×750, 4×150, 5×75, 6×50 in less than half an hour, and of sizes 4×200, 5×100 in less than an hour.

In constructing the bounded regions we have tried the different variants proposed in Sect. 11.3. As it was observed, all proposed approaches (B, C, D) that use the subgradient's ending point or direction to define the next cutting point converge extremely slow, compared to the simple divisive approach (A). Although the subgradients may sometimes produce really good cutting points, which strongly reduce the bounded region, they often stick at the angles of the bounded region, so that the next steps reduce the bounded region by a narrow slice only, which is close to an existing bound. Particularly in higher dimensions, the subgradients also appear to be too short, so that the amount of the volume cut in each step becomes unsatisfying. Therefore in out search we desist from the lengths of the subgradients and use only their directions. All the numerical results provided in this paper were received using the divisive approach starting with the initial rectangular bounded region. The number of cuts needed to obtain the desired volume appears to depend only moderately on the size and dimensionality of the data.

For both algorithms, the ROO and the new one, the performance of the searching procedure strictly depends on the selected initial line. As ROO select this line at random, their calculation times differ significantly between different launches. Our bounding algorithm selects the firstly found border line of the bounded region as the initial line, which makes the searching path completely deterministic, although in general not the fastest possible.

Employing bounds considerably decreases the complexity of the algorithm. The minimization along a line produces most of the complexity of the ROO algorithm. The line is intersected with $H = \binom{n}{k}$ hyperplanes, the intersections are sorted and all of them iterated, which has a complexity of $O(H^2 \log H)$. The bounding algorithm leaves a smaller amount $h < H$ of hyperplanes. Only b hyperplanes, $b < h$, that have intersections between the bounds remain to be considered. The rougher bound also strongly decreases the number of hyperplanes which need to be sorted to s : $b < s < h$. This provides a complexity of $O(b \times h \log s)$ only.

Tables 11.1, 11.2, and 11.3 exhibit a few exemplary results. The experimental data is an even mixture of two multidimensional normal distributions $N(\mathbf{0}_k, diag(\mathbf{1}_k))$ and $N([15, 0, \ldots, 0]_k, diag([1, 25, 1, \ldots, 1]_k))$ although the conclusions are the same for the data having other form and for the real data sets. They show how the performance parameters listed below depend on the data dimension and size, given the intended volume equal to 10^{-8} (for Tables 11.1 and 11.2), where *#Cuts* is the number of cuts needed to reach the intended volume using the divisive approach, *H. planes (%)* is the percent of the hyperplanes intersecting the bounded region, *#Steps* is the number of minimization steps needed to find the median, and time periods needed to: determine the bounded region T_{bounds}, calculate the median (after the bounded region is determined) T_{count}, perform the whole procedure T_{total}, and to find the median using the algorithm by ROO $T_{original}$. The given times do not include the generation of all observation hyperplanes, which is the same for both algorithms.

The part of the hyperplanes crossing the bounded region of the given volume grows quickly with dimension, as it is seen in Tables 11.1 and 11.3. On the other hand, the part of these hyperplanes that take part in the minimization process decreases, since many of their intersections with a searching line lie outside the bounded region. Note that the bounded region, being located in the middle of

Table 11.1 The performance parameters for $n \in \{50, 75\}$ and intended volume 10^{-8}

k	n	#Cuts	H. planes (%)	#Steps	T_{bounds}	T_{count}	T_{total}	$T_{original}$
2	50	29	0.16	3	0.009	0.001	0.010	0.018
3	50	39	0.64	9	0.216	0.053	0.269	0.557
4	50	42	3.33	34	3.015	2.433	5.448	25.529
5	50	42	9.65	45	31.359	42.876	74.235	476.600
6	50	45	17.65	77	345.128	774.382	1119.510	3149.010
2	75	32	0.11	2	0.023	0.002	0.025	0.038
3	75	36	0.34	14	0.658	0.243	0.901	3.033
4	75	42	2.11	39	15.353	13.720	29.073	110.291
5	75	45	7.17	70	281.888	474.461	756.349	2667.890

Table 11.2 The performance parameters for $k \in \{4, 5\}$ and intended volume 10^{-8}

k	n	#Cuts	H. planes (%)	#Steps	T_{bounds}	T_{count}	T_{total}	$T_{original}$
4	25	38	3.26	25	0.159	0.095	0.254	0.707
4	50	42	3.33	34	3.015	2.433	5.448	25.529
4	75	42	2.11	39	15.353	13.720	29.073	110.291
4	100	43	2.60	35	49.360	41.691	91.051	338.950
5	25	44	11.77	39	0.930	0.932	1.862	6.171
5	50	42	9.65	45	31.359	42.876	74.235	476.600
5	75	45	7.17	70	281.888	474.461	756.349	2667.890
5	100	43	8.53	71	1166.930	2220.330	3387.260	9803.730

Table 11.3 The performance parameters for data sets 4×100 and 6×50, with different intended volumes

k	n	Volume	#Cuts	H. planes (%)	#Steps	T_{bounds}	T_{count}	T_{total}
4	100	1			36			338.950
4	100	10^{-02}	14	67.45	50	17.648	283.936	301.584
4	100	10^{-03}	18	46.62	49	23.216	201.245	224.461
4	100	10^{-04}	22	28.47	39	28.304	112.429	140.733
4	100	10^{-05}	27	14.23	48	33.403	97.364	130.767
4	100	10^{-06}	31	9.76	29	37.707	49.499	87.206
4	100	10^{-07}	37	5.47	38	44.391	55.032	99.423
4	100	10^{-08}	43	2.60	35	49.360	41.691	91.051
4	100	10^{-09}	47	1.64	33	57.209	37.402	94.611
4	100	10^{-10}	52	0.83	33	59.709	35.347	95.056
4	100	10^{-20}	97	<0.01	22	108.205	21.089	129.294
4	100	10^{-30}	100	<0.01	13	109.004	12.778	121.782
6	50	1			73			3149.010
6	50	10^{-05}	31	53.86	91	215.747	1842.733	2058.480
6	50	10^{-06}	36	35.61	71	270.619	1272.041	1542.660
6	50	10^{-07}	40	26.68	75	301.413	885.707	1187.120
6	50	10^{-08}	45	17.65	77	345.128	774.382	1119.510
6	50	10^{-09}	49	12.97	70	373.170	545.055	918.225
6	50	10^{-10}	53	9.14	94	378.424	705.386	1083.810

Volume equal to one corresponds to the ROO algorithm

the data cloud, is intersected by most of the hyperplanes, so that the part of the included hyperplanes is much larger than the part of the final volume, compared to the initial one. Our calculations demonstrate that the part of included hyperplanes strongly depends on the dimensionality and the number of observations, which is also shown in Fig. 11.6. However, the number of observations has less influence than the dimension.

These three tables also show that the new exact bounding algorithm finds the median much faster than the one of ROO. We observe that for each given data set the number of necessary minimization steps is almost the same in both algorithms. As the intended volume is reduced, the time needed to build the bounded region increases, while the minimization time decreases along with the number of hyperplanes and their intersections involved, and the total time also decreases (Table 11.3, Fig. 11.7). However, beyond some point, usually at around 10^{-08} of the volume, this procedure becomes less efficient, and the total time increases. A smaller volume may also contain a higher amount of isolated routes through the observation lines, which involves travelling along the bounds and additionally slows the procedure down.

If the volume is small enough, any point of it (e.g., the average of the bounds' intersections) may be taken as an approximate value of a median. For example, for a four-dimensional dataset bounded by a cube of side length 10, 10^{-8} of the

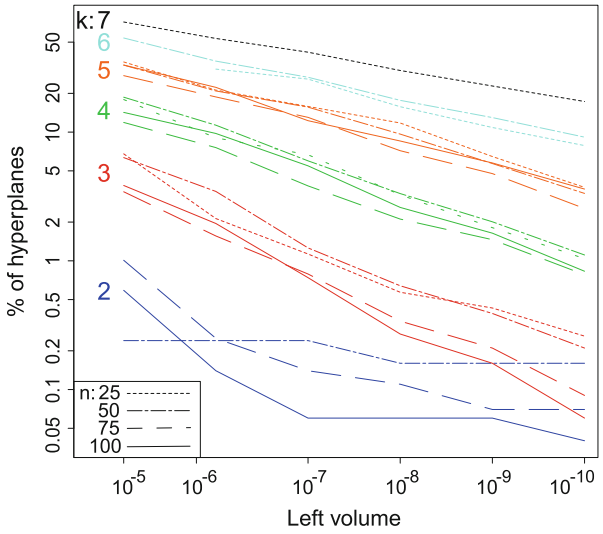

Fig. 11.6 The dependence of the part of hyperplanes crossing the bounded region (log scale) on the size of the region for $k \in [2..7]$ and $n \in \{25, 50, 75, 100\}$

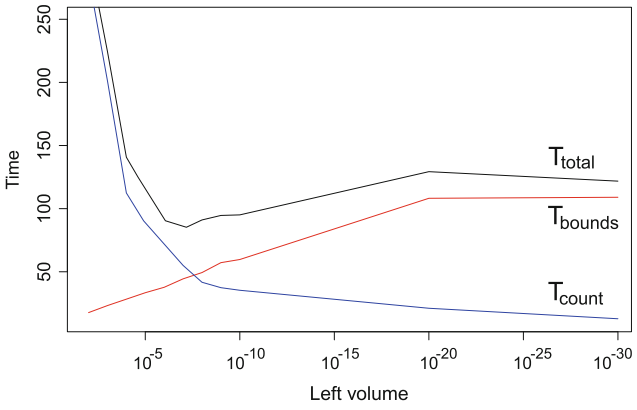

Fig. 11.7 The dependence of calculation time on the size of the bounded region. Note that the ROO algorithm has total time of ca. 340 s

volume was reached in 43 cuts, and the center of the final bounded region equalled the median ± 0.05 by each coordinate, which is quite precise. The precision of this approximative method depends on the volume of the bounded region and is controlled by it. The method yields as precise results as the approximative methods provided in the *OjaNP* R-package. In general, the approximate value of the median is found much faster than the exact one. We searched an approximative median with precision equal to half of the bounded region's volume, the computation times of which are given in the column T_{bounds} of Table 11.3. However, this approximation

method is not really useful, as it considers all observation hyperplanes and is therefore largely outperformed by the approximative methods of ROO.

References

Fischer, D., Möttönen, J., Nordhausen, K., Vogel, D.: OjaNP: Multivariate methods based on the Oja median and related concepts. R package version 0.9-8 (2014). http://cran.r-project.org/web/packages/OjaNP/

Hettmansperger, T.P., Möttönen, J., Oja, H.: The geometry of the affine invariant multivariate sign and rank methods. J. Nonparametr. Stat. **11**, 271–285 (1999)

Koshevoy, G.: Lift-zonoid and multivariate depths. In: Dutter, R. (ed.) Developments in Robust Statistics, pp. 194–202. Springer, Berlin (2003)

Liu, R.Y.: On a notion of simplicial depth. Proc. Natl. Acad. Sci. **85**, 1732–1734 (1988)

Mosler, K.: Depth statistics. In: Becker, C., Fried, R., Kuhnt, S. (eds.) Robustness and Complex Data Structures. Festschrift in Honour of Ursula Gather, pp. 17–34. Springer, Berlin (2013)

Niinimaa, A., Oja, H., Tableman, M.: The finite-sample breakdown point of the Oja bivariate median and of the corresponding half-samples version. Stat. Probab. Lett. **10**, 325–328 (1990)

Niinimaa, A., Oja, H., Nyblom, J.: Algorithm AS 277. The Oja bivariate median. Appl. Stat. **41**, 611–617 (1992)

Oja, H.: Descriptive statistics for multivariate distributions. Stat. Probab. Lett. **1**, 327–332 (1983)

Oja, H.: Multivariate median. In: Becker, C., Fried, R:, S. Kuhnt, S. (eds.) Robustness and Complex Data Structures. Festschrift in Honour of Ursula Gather, pp. 3–15. Springer, Berlin (2013)

Ronkainen, T., Oja, H., Orponen, P.: Computation of the multivariate Oja median. In: Dutter, R. (ed.) Developments in Robust Statistics, pp. 344–359. Springer, Heidelberg (2003)

Small, C.G.: Multidimensional medians arising from goedesics on graphs. Ann. Stat. **25**, 478–494 (1997)

Tukey, J.W.: Mathematics and picturing data. In: James, R.D. (ed.) Proceedings of the 1974 International Congress of Mathematicians, Vancouver, pp. 523–531 (1975)

Zuo, Y., Serfling, R.: General notions of statistical depth function. Ann. Stat. **28**, 461–482 (2000)

Chapter 12
Algorithms for the Spatial Median

John T. Kent, Fikret Er, and Patrick D.L. Constable

Abstract The spatial median can be defined as the unique minimum of a strictly convex objective function. Hence, its computation through an iterative algorithm ought to be straightforward. The simplest algorithm is the steepest descent Weiszfeld algorithm, as modified by Ostresh and by Vardi and Zhang. Another natural algorithm is Newton-Raphson. Unfortunately, all these algorithms can have problems near data points; indeed, Newton-Raphson can converge to a non-optimal data point, even if a line search is included! However, by combining these algorithms, a reliable and efficient "hybrid" algorithm can be developed.

Keywords EM algorithm • MM algorithm • Newton-Raphson algorithm • Steepest descent • Vardi-Zhang algorithm • Weiszfeld algorithm

12.1 Introduction

Given n observations $\mathbf{x}_i \in \mathbb{R}^p$ and fixed weights $w_i > 0$, $i = 1, \ldots, n$, the spatial median $\hat{\boldsymbol{\mu}} \in \mathbb{R}^p$ is defined to minimize

$$g(\boldsymbol{\mu}) = \sum_{i=1}^{n} g_i(\boldsymbol{\mu}) = \sum_{i=1}^{n} w_i |\boldsymbol{\mu} - \mathbf{x}_i| \tag{12.1}$$

where $|\boldsymbol{\mu} - \mathbf{x}_i| = \{\sum_{j=1}^{p} (\mu_j - x_{ij})^2\}^{1/2}$ denotes the usual Euclidean distance between $\boldsymbol{\mu}$ and $\mathbf{x}_i$. The spatial median has a long history in Statistics as a robust alternative to

J.T. Kent (✉)
Department of Statistics, University of Leeds, Leeds LS2 9JT, UK
e-mail: j.t.kent@leeds.ac.uk

F. Er
Open Education Faculty, Yunusemre Campus, Anadolu University, 26470 Eskisehir, Turkey
e-mail: fer@anadolu.edu.tr

P.D.L. Constable
Stony Lodge, Holwell, Sherborne, Dorset DT9 5LQ, UK
e-mail: PatCnstb@aol.com

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_12

the sample mean vector. Kuhn (1973) traces the idea back to Fermat in the 1600s and Simpson in the 1700s. Other reviews include Small (1990), Vardi and Zhang (2000), and Oja (2010). Some of the more mainstream statistical contributions include Haldane (1948) and Gower (1974). Brown (1983) discusses issues of statistical inference. Other names for the spatial median are the "median center" and the "multivariate L_1 median."

The purpose of this paper is to look at the numerical problem of iteratively computing $\hat{\boldsymbol{\mu}}$. Surprisingly, given all the attention that this problem has received over the years, it still seems possible to give some new insights.

Provided the $\{\mathbf{x}_i\}$ are not collinear, $g(\boldsymbol{\mu})$ is a strictly convex function with a unique minimum $\hat{\boldsymbol{\mu}}$. As a first attempt to compute the spatial median, it seems sensible to try standard optimization methods based on first and second derivatives. However, there are two complications: (1) $\hat{\boldsymbol{\mu}}$ may sometimes coincide with one of the data points $\{\mathbf{x}_i\}$ (a minor issue), and (2) the derivative of g does not exist at the data points (a more major issue).

Differentiating (12.1) and setting the gradient to 0 yields the equation

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{n} w_i \mathbf{x}_i / d_i}{\sum_{i=1}^{n} w_i / d_i}, \quad d_i = |\hat{\boldsymbol{\mu}} - \mathbf{x}_i|, \tag{12.2}$$

which must be satisfied if $\hat{\boldsymbol{\mu}}$ does not coincide with a data point. This identity suggests an iterative "reweighting algorithm" based on the following updating function. Let

$$\mathbf{M}(\boldsymbol{\mu}) = \frac{\sum_{i=1}^{n} w_i \mathbf{x}_i / d_i}{\sum_{i=1}^{n} w_i / d_i}, \quad d_i = |\boldsymbol{\mu} - \mathbf{x}_i|, \tag{12.3}$$

which is well defined provided $\boldsymbol{\mu}$ does not coincide with any of the data points. Thus $\mathbf{M}(\boldsymbol{\mu})$ is an adaptive weighted average of the data where the weights w_i/d_i depend on $\boldsymbol{\mu}$. This reweighting algorithm is often named after Weiszfeld (see Kuhn 1973), who seems to have been to first to discover it.

Note that if $\boldsymbol{\mu}$ is close to one of the data points, $\mathbf{x}_k$, say, then $d_k \to 0$ whereas all the d_i, $i \neq k$, remain bounded away from 0; hence, $\mathbf{M}(\boldsymbol{\mu}) \to \mathbf{x}_k$ as $\boldsymbol{\mu} \to \mathbf{x}_k$. Thus we extend the definition of $\mathbf{M}(\boldsymbol{\mu})$ by continuity to include the data points, $\mathbf{M}(\mathbf{x}_i) = \mathbf{x}_i$, $i = 1, \ldots, n$; that is, the data points are fixed points of the Weiszfeld algorithm.

Ostresh (1978) suggested a modified reweighting algorithm for which the only fixed point is $\hat{\boldsymbol{\mu}}$. It was independently rediscovered by Vardi and Zhang (2000), who gave some additional properties. We shall label this algorithm the WOVZ algorithm. Let $\mathbf{M}^*(\boldsymbol{\mu})$ denote the WOVZ update function. Details are given in Sect. 12.5. It can be shown that the WOVZ update is always a steepest descent update. Further, given any initial estimate $\boldsymbol{\mu}^{(0)}$, the iterates defined by $\boldsymbol{\mu}^{(\nu+1)} = \mathbf{M}^*(\boldsymbol{\mu}^{(\nu)})$ will always converge to $\hat{\boldsymbol{\mu}}$.

Thus here are some possible strategies for optimization, though problems can be expected with derivative-based methods at or near data points.

1. Enumeration. Check some or all the data points to see if they minimize g.
2. Derivative-free minimization methods such as the simplex or Nelder–Mead algorithm. Such methods are reliable, but they are slower than suitably formulated derivative-based methods, and are not considered further here.
3. First derivative (steepest descent) minimization methods, especially the WOVZ algorithm, possibly with line search.
4. Second derivative minimization methods such as Newton-Raphson (NR), possibly with line search.
5. More general convex optimization methods. These are discussed briefly in Sect. 12.11.

There are two main objectives of this paper. The first is to give a detailed study of the Weiszfeld and WOVZ algorithms. After setting out some basic properties of $g(\boldsymbol{\mu})$ in Sect. 12.2, we analyze the behavior of Weiszfeld update $\mathbf{M}(\boldsymbol{\mu})$ near the data points in Sect. 12.3. Both EM and MM interpretations for the Weiszfeld algorithm are given in Sect. 12.4, each of which ensures monotonicity of the algorithms. However, the Weiszfeld algorithm is a very unusual EM algorithm because all the data points are fixed points. This "flaw" is fixed by the WOVZ algorithm, whose properties are examined in Sect. 12.5. The special one-dimensional case $p = 1$ is studied in Sect. 12.6.

The Newton-Raphson algorithm is studied in Sect. 12.7. It has the unusual property in this setting that, even with step-halving, it can converge to a non-optimal data point.

Thus neither WOVZ nor NR is a very good algorithm on its own to compute the spatial median. However, the second main objective of the paper is to show that by combining WOVZ and NR together into a hybrid algorithm, it is possible to obtain a reliable and efficient algorithm. Here reliability means that the algorithm should never get stuck at a non-optimal data point. Efficiency means that the number of iterations should not become indefinitely large for a starting point arbitrarily close to a data point. The need to state these requirements explicitly will become clear when we study various algorithms in detail.

Our final recommendation is a hybrid algorithm, where at each iteration we choose the best possibility from

1. The nearest data point to the current estimate of $\boldsymbol{\mu}$,
2. The WOVZ update, and
3. The NR update, with step-halving.

The rationale behind this choice is discussed in detail in Sect. 12.8 with some numerical examples in Sect. 12.9. The paper finishes with a brief discussion of existing software in `R` (Sect. 12.10) and more general convex optimization algorithms (Sect. 12.11), together with a recap of the main conclusions (Sect. 12.12).

Hannu Oja has been one of the leading researchers into nonparametric and robust methods for multivariate data. It is a great pleasure to include this paper in a volume for his 65th birthday.

12.2 Basic Properties of $g(\boldsymbol{\mu})$

Except for Sect. 12.6 we shall always assume that $p \geq 2$ and that the data points $\{\mathbf{x}_i\}$ are distinct and do not lie on a single line in $\mathbb{R}^p$. The first two derivatives of $g(\boldsymbol{\mu})$ are given by

$$\nabla g(\boldsymbol{\mu}) = \sum w_i(\boldsymbol{\mu} - \mathbf{x}_i)/d_i, \quad d_i = |\boldsymbol{\mu} - \mathbf{x}_i|, \tag{12.4}$$

$$\nabla\nabla^T g(\boldsymbol{\mu}) = \sum w_i\{\mathbf{I}/d_i - (\boldsymbol{\mu} - \mathbf{x}_i)(\boldsymbol{\mu} - \mathbf{x}_i)^T/d_i^3\}, \tag{12.5}$$

except at the data points where the derivatives do not exist. The ith term of (12.5) is a positive semi-definite matrix of rank $p-1$. Since the $\{(\boldsymbol{\mu} - \mathbf{x}_i)\}$ cannot all lie in a common one-dimensional subspace, the terms of (12.5) cannot have the same nullspace. Hence $\nabla\nabla^T g(\boldsymbol{\mu})$ is positive definite, except at the data points where it does not exist.

It is also useful to define a "restricted" objective function

$$g_{\backslash k}(\boldsymbol{\mu}) = \sum_{i \neq k} w_i|\boldsymbol{\mu} - \mathbf{x}_i|, \tag{12.6}$$

by excluding the kth data point. Although the full objective function is not differentiable at $\boldsymbol{\mu} = \mathbf{x}_k$, the restricted objective function is. Set

$$\mathbf{v}_k = \nabla g_{\backslash k}(\mathbf{x}_k) = \sum_{i \neq k} w_i(\mathbf{x}_k - \mathbf{x}_i)/d_{ik}, \quad d_{ik} = |\mathbf{x}_k - \mathbf{x}_i| \tag{12.7}$$

and define the "repulsiveness" parameter for $\mathbf{x}_k$

$$\alpha_k = |\nabla g_{\backslash k}(\mathbf{x}_k)|/w_k = |\mathbf{v}_k|/w_k. \tag{12.8}$$

It is well known that $g(\boldsymbol{\mu})$ is strictly convex. Hence the minimizer $\hat{\boldsymbol{\mu}}$ of (12.1) is unique. Further, depending on the layout of the data, $\hat{\boldsymbol{\mu}}$ may or may not coincide with one of the data points. In order to investigate this possibility, we look at the behavior of g and its directional derivatives near a data point, $\mathbf{x}_k$, say. Let $\mathbf{e}_0 \in \mathbb{R}^p$ be a unit vector and let $\boldsymbol{\mu} = \mathbf{x}_k + \varepsilon\mathbf{e}_0$ where $\varepsilon > 0$ is small. By Taylor's theorem

$$\begin{aligned} g(\boldsymbol{\mu}) &= w_k\varepsilon + g_{\backslash k}(\mathbf{x}_k + \varepsilon\mathbf{e}_0) \\ &= w_k\varepsilon + g_{\backslash k}(\mathbf{x}_k) + \varepsilon\mathbf{e}_0^T\mathbf{v}_k + \frac{1}{2}\varepsilon^2\mathbf{e}_0^T\{\nabla\nabla^T g_{\backslash k}(\mathbf{x}_k)\}\mathbf{e}_0 + O(\varepsilon^3) \end{aligned} \tag{12.9}$$

where $\mathbf{v}_k = \nabla g_{\backslash k}(\mathbf{x}_k)$ and $\nabla\nabla^T g_{\backslash k}(\mathbf{x}_k)$ are well defined at $\boldsymbol{\mu} = \mathbf{x}_k$, and where the $O(\varepsilon^3)$ term is bounded uniformly over the directions $\mathbf{e}_0$.

The non-collinearity of the data ensures that the span of the directions $\{(\mathbf{x}_k - \mathbf{x}_i) : i = 1, \ldots, n, \ i \neq k\}$ is $\mathbb{R}^p$, so that the second derivative matrix of the

restricted objective function, $\nabla\nabla^T g_{\setminus k}(\mathbf{x}_k)$, is positive definite. Further, although the full objective function is not differentiable at $\boldsymbol{\mu} = \mathbf{x}_k$, its directional derivatives are well defined. Letting $\epsilon \to 0$ in (12.9) we find the partial derivative $\partial_{\mathbf{e}_0} g(\mathbf{x}_k)$ is given by

$$\partial_{\mathbf{e}_0} g(\mathbf{x}_k) = w_k + \mathbf{e}_0^T \mathbf{v}_k. \tag{12.10}$$

We can now provide a simple condition based on the repulsiveness parameters to check whether $\hat{\boldsymbol{\mu}} = \mathbf{x}_k$ for any choice of k.

Proposition 12.1 *The spatial median satisfies $\hat{\boldsymbol{\mu}} = \mathbf{x}_k$ for some k, $1 \le k \le n$, if and only if $\alpha_k \le 1$, a condition which can hold for at most one index k.*

Proof If $\alpha_k < 1$ in (12.8), then $-\mathbf{e}_0^T \mathbf{v}_k \le |\mathbf{v}_k| = \alpha_k w_k < w_k$ for all unit vectors $\mathbf{e}_0$. Hence the directional derivatives at $\boldsymbol{\mu} = \mathbf{x}_k$ satisfy

$$\partial_{\mathbf{e}_0} g(\mathbf{x}_k) = w_k + \mathbf{e}_0^T \mathbf{v}_k > 0$$

for all $\mathbf{e}_0$. That is, $\boldsymbol{\mu} = \mathbf{x}_k$ is a local minimum of g, and hence the global minimum.

If $\alpha_k = 1$, then $-\mathbf{e}_0^T \mathbf{v}_k < w_k$ for all $\mathbf{e}_0$ except $\mathbf{e}_0 = -\mathbf{v}_k/(\alpha_k w_k)$. In this direction it is necessary to go to the second term in the Taylor series expansion to ensure that $\boldsymbol{\mu} = \mathbf{x}_k$ is a local (and hence the global) minimum.

Finally, if $\alpha_k > 1$, then $g(\boldsymbol{\mu})$ decreases in the direction $\mathbf{e}_0 = -\mathbf{v}_k/(\alpha_k w_k)$, so that $\boldsymbol{\mu} = \mathbf{x}_k$ cannot be the minimizing value.

In the language of convex optimization, a subgradient of $g(\mathbf{x}_k)$ is a vector $\mathbf{v}$ such that $g(\mathbf{x}) - g(\mathbf{x}_k) \ge \mathbf{v}^T(\mathbf{x} - \mathbf{x}_k)$ for all $\mathbf{x}$. The set of all subgradients at $\mathbf{x}_k$ is called the subdifferential at $\mathbf{x}_k$ and is denoted $\partial g(\mathbf{x}_k)$. Then the proposition can be recast to state that $\alpha_k \le 1$ if and only if the origin is in the subdifferential $\partial g(\mathbf{x}_k)$.

As a result of the proposition, let us say that $\mathbf{x}_k$ is "sticky" if $\alpha_k \le 1$ and "nonsticky" if $\alpha_k > 1$.

12.3 Behavior of the Weiszfeld Algorithm

In this section we summarize some of the basic properties of the Weiszfeld algorithm.

12.3.1 Steepest Descent Property

Suppose $\boldsymbol{\mu}$ is not equal to any of the data points. From (12.3) and (12.4) with $d_i = |\boldsymbol{\mu} - \mathbf{x}_i|$,

$$\begin{aligned} \mathbf{M}(\boldsymbol{\mu}) - \boldsymbol{\mu} &= \frac{\sum w_i(\mathbf{x}_i - \boldsymbol{\mu})/d_i}{\sum w_i/d_i} \\ &\propto -\nabla g(\boldsymbol{\mu}), \end{aligned} \tag{12.11}$$

so that $\mathbf{M}(\boldsymbol{\mu})$ lies on the line of steepest descent from $\boldsymbol{\mu}$. If $\boldsymbol{\mu}$ equals one of the data points, then $\boldsymbol{\mu}$ is a fixed point of the algorithm, $\mathbf{M}(\boldsymbol{\mu}) = \boldsymbol{\mu}$, so the algorithm still has the steepest descent property, at least in a degenerate sense.

12.3.2 Behavior Near Data Points

Let $\boldsymbol{\mu} = \mathbf{x}_k + \varepsilon \mathbf{e}_0$ lie near a data point. Kuhn (1973) gives an expansion for the updating function. From (12.3),

$$\begin{aligned} \mathbf{M}(\boldsymbol{\mu}) - \mathbf{x}_k &= \frac{\sum w_i(\mathbf{x}_i - \mathbf{x}_k)/d_i}{\sum w_i/d_i} \\ &= \frac{\mathbf{0} + \sum_{i \neq k} w_i(\mathbf{x}_i - \mathbf{x}_k)/d_i}{w_k/\varepsilon + \sum_{i \neq k} w_i/d_i} \\ &= -\frac{\varepsilon}{w_k} \sum_{i \neq k} w_i(\mathbf{x}_k - \mathbf{x}_i)/d_{ik} + \mathbf{O}(\varepsilon^2) \\ &= -\frac{\varepsilon}{w_k}\mathbf{v}_k + \mathbf{O}(\varepsilon^2). \end{aligned} \tag{12.12}$$

Note that in the last two lines we have replaced $d_i = |\boldsymbol{\mu} - \mathbf{x}_i|$ by $d_{ik} = |\mathbf{x}_k - \mathbf{x}_i|$ for $i \neq k$, with the difference absorbed in the $\mathbf{O}(\varepsilon^2)$ term. There are several important messages from (12.12).

1. To first order $\mathbf{M}(\boldsymbol{\mu})$ does not depend on $\mathbf{e}_0$ but lies on the ray of steepest descent for the restricted objective function $g_{\setminus k}(\boldsymbol{\mu})$ at $\boldsymbol{\mu} = \mathbf{x}_k$. If $\mathbf{x}_k$ is nonsticky ($\alpha_k > 1$), then from (12.10), this ray is also the ray of steepest descent for $g(\boldsymbol{\mu})$ at $\boldsymbol{\mu} = \mathbf{x}_k$. On the other hand, if $\mathbf{x}_k$ is sticky ($\alpha_k < 1$), this direction is the ray of shallowest ascent for $g(\boldsymbol{\mu})$ at $\boldsymbol{\mu} = \mathbf{x}_k$.
2. $|\mathbf{M}(\boldsymbol{\mu}) - \mathbf{x}_k)| \cong \alpha_k |\boldsymbol{\mu} - \mathbf{x}_k|$. Thus, if $\alpha_k < 1$ ($\alpha_k > 1$), the Weiszfeld algorithm converges to (diverges from) $\mathbf{x}_k$ at linear rate α_k along this steepest descent (shallowest ascent) direction to (from) $\mathbf{x}_k$.

3. In particular, if α_k is equal or near to 1, the movement of the updating algorithm can be very slow.

12.4 EM and MM Interpretations for the Updating Algorithm

Suppose the data $\{\mathbf{x}_i\}$ are realizations of random vectors $\{\mathbf{X}_i\}$ of the form $\mathbf{X}_i = \boldsymbol{\mu} + \mathbf{U}_i/\sqrt{w_i}$, where the $\{\mathbf{U}_i\}$ are i.i.d from a pdf

$$f(\mathbf{u}) \propto \exp(-\rho(s)), \quad s = \mathbf{u}^T\mathbf{u}, \quad \mathbf{u} \in \mathbb{R}^p, \tag{12.13}$$

where $\rho(s)$ is a continuous function of $s \geq 0$. Up to a constant term, minus the log-likelihood takes the form

$$g(\boldsymbol{\mu}) = \sum_{i=1}^{n} w_i \rho(s_i), \quad s_i = (\mathbf{x}_i - \boldsymbol{\mu})^T(\mathbf{x}_i - \boldsymbol{\mu}). \tag{12.14}$$

The spatial median coincides with the mle of $\boldsymbol{\mu}$ for this weighted location problem under the choice $\rho(s) = \sqrt{s},\ s \geq 0$.

In general, if $\exp\{-\rho(s)\}$ is completely monotone (that is, the derivatives of $\exp(-\rho(s))$ satisfy $(-1)^m(d/ds)^m \exp\{-\rho(s)\} > 0$ for $m \geq 0, s > 0$), then the $\{\mathbf{U}_i\}$ can be represented as scale mixtures of isotropic normal densities,

$$\mathbf{U}_i = \mathbf{Z}_i/\sqrt{Y_i}, \quad \mathbf{Z}_i \sim N_p(\mathbf{0}, \mathbf{I}_p), \tag{12.15}$$

where the distribution of Y_i on $(0, \infty)$ is determined by $\rho(\cdot)$ (Andrews and Mallows 1974). In this case, the EM algorithm can be used to iteratively compute the mle for $\boldsymbol{\mu}$, treating the $\{Y_i\}$ as missing data (Dempster et al. 1977, 1980; McLachlan and Krishnan 2008). The updating function takes the form

$$\mathbf{M}^{(\mathrm{EM})}(\boldsymbol{\mu}) = \sum \hat{y}_i w_i \mathbf{x}_i / \sum \hat{y}_i w_i, \quad \hat{y}_i = \rho'(s_i), \quad s_i = |\mathbf{x}_i - \boldsymbol{\mu}|^2. \tag{12.16}$$

A sufficient condition for $\exp\{-\rho(s)\}$ to be completely monotone is that $\rho'(s) > 0$ for $s > 0$ with completely monotone derivative (Feller 1966, p. 417). Clearly, $\rho(s) = \sqrt{s}$ has this property, and (12.16) reduces to the Weiszfeld update (12.3). Hence the Weiszfeld algorithm has an EM interpretation. Several consequences can be drawn from this identification.

1. It is a general property of EM algorithms that, except at fixed points of the algorithm, the likelihood is strictly increased at each iteration. In our notation

$$g(\mathbf{M}(\boldsymbol{\mu})) < g(\boldsymbol{\mu}) \text{ unless } \boldsymbol{\mu} \text{ is a fixed point.} \tag{12.17}$$

2. In regular EM problems the fixed points of the EM algorithm coincide with the stationary points of the likelihood (Wu 1983). Since $g(\boldsymbol{\mu})$ in (12.1) is strictly convex, it has a unique stationary point which is its global minimum. However, $\rho(s) = \sqrt{s}$ is not differentiable at $s = 0$, which means that the corresponding EM problem is not regular. This lack of regularity explains how the Weiszfeld algorithm can also have fixed points at each of the data points. Thus, the Weiszfeld algorithm provides a simple example of a nonregular EM algorithm with additional stationary points.

The EM algorithm for the location problem from a scale mixture of normal distributions also arises as an MM algorithm (for majorize-minimize). The MM assumptions are more general and merely require that $\rho(s)$ be continuous, increasing and concave for $s \geq 0$ and continuously differentiable for $s > 0$. (If $\exp\{-\rho(s)\}$ is completely monotone, the concavity is automatic for $\rho(s)$, and in particular these conditions are satisfied by $\rho(s) = \sqrt{s}$.) The basic idea of an MM algorithm is to find for any $\boldsymbol{\mu}_0$ a "majorizing function" $G(\boldsymbol{\mu}, \boldsymbol{\mu}_0)$ such that

(a) $g(\boldsymbol{\mu}) \leq G(\boldsymbol{\mu}, \boldsymbol{\mu}_0)$ for all $\boldsymbol{\mu}$, and
(b) $g(\boldsymbol{\mu}_0) = G(\boldsymbol{\mu}_0, \boldsymbol{\mu}_0)$.

If $G(\boldsymbol{\mu}, \boldsymbol{\mu}_0)$ can be minimized explicitly over $\boldsymbol{\mu}$ to yield $\boldsymbol{\mu}_1 = \mathbf{M}^{(\mathrm{MM})}(\boldsymbol{\mu}_0)$, say, then necessarily $g(\boldsymbol{\mu}_1) \leq g(\boldsymbol{\mu}_0)$, usually with strict inequality.

When $g(\boldsymbol{\mu})$ takes the form (12.14), a convenient majorizing function is given by

$$G(\boldsymbol{\mu}, \boldsymbol{\mu}_0) = g(\boldsymbol{\mu}_0) + \sum w_i (s_i - s_i^{(0)}) \rho'(s_i^0), \tag{12.18}$$

where $s_i = |\mathbf{x}_i - \boldsymbol{\mu}|^2$, $s_i^{(0)} = |\mathbf{x}_i - \boldsymbol{\mu}_0|^2$. The majorizing property follows from the concavity of $\rho(s)$. More specifically, expand $\rho(s_i)$ to first-order in Taylor series in s_i at $s_i = s_i^0$ and note that the remainder term is nonpositive. Further, provided $\boldsymbol{\mu}_0$ is distinct from all the data points, $G(\boldsymbol{\mu}, \boldsymbol{\mu}_0)$ is well defined and minimizing $G(\boldsymbol{\mu}, \boldsymbol{\mu}_0)$ over $\boldsymbol{\mu}$ is a weighted least squares problem leading to the standard update (12.16).

MM arguments have been used by many authors, e.g. Kuhn (1973) for the spatial median, Huber (1981) for M-estimators of location, and Lange and Sinsheimer (1993) for the setting given here. The methodology can also cope with the inclusion of a scale parameter (not needed for the spatial median). A more general version of the MM methodology forms the basis of a broad theory of optimization; see Lange et al. (2000) and Hunter and Lange (2004).

To round off the discussion, we use the MM framework to show that for the spatial median problem there are no other fixed points of the Weiszfeld algorithm besides the data points and $\hat{\boldsymbol{\mu}}$.

Proposition 12.2 *For the spatial median problem, if* $\boldsymbol{\mu}_0 \notin \{\hat{\boldsymbol{\mu}}\} \bigcup \{\mathbf{x}_i\}_1^n$ *then* $g(\mathbf{M}(\boldsymbol{\mu}_0)) < g(\boldsymbol{\mu}_0)$.

Proof Suppose $\boldsymbol{\mu}_0 \notin \{\hat{\boldsymbol{\mu}}\} \bigcup \{\mathbf{x}_i\}_1^n$. Then the first derivatives in (12.18) are well defined and finite. The weighted least squares criterion in (12.18) (with $\boldsymbol{\mu}_0$ held

fixed) has a unique minimum for $\boldsymbol{\mu}$. Hence if $\mathbf{M}(\boldsymbol{\mu}_0) \neq \boldsymbol{\mu}_0$, it follows that $G(\mathbf{M}(\boldsymbol{\mu}_0), \boldsymbol{\mu}_0) < G(\boldsymbol{\mu}_0, \boldsymbol{\mu}_0)$ and so $g(\mathbf{M}(\boldsymbol{\mu}_0)) < g(\boldsymbol{\mu}_0)$.

It remains to be checked that $\mathbf{M}(\boldsymbol{\mu}_0) \neq \boldsymbol{\mu}_0$. Note that $\nabla g(\boldsymbol{\mu}) = \nabla G(\boldsymbol{\mu}, \boldsymbol{\mu}_0)$ at $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ (treating $G(\boldsymbol{\mu}, \boldsymbol{\mu}_0)$ as a function of $\boldsymbol{\mu}$ with $\boldsymbol{\mu}_0$ held fixed). If $\boldsymbol{\mu}_0 \neq \hat{\boldsymbol{\mu}}$, then $\nabla g(\boldsymbol{\mu}_0) \neq 0$, so that $\nabla G(\boldsymbol{\mu}, \boldsymbol{\mu}_0) \neq 0$ at $\boldsymbol{\mu} = \boldsymbol{\mu}_0$; hence, $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ does not minimize $G(\boldsymbol{\mu}, \boldsymbol{\mu}_0)$, so that $\mathbf{M}(\boldsymbol{\mu}_0) \neq \boldsymbol{\mu}_0$.

12.5 The WOVZ Algorithm

Once again we focus on the spatial median problem with $\rho(s) = \sqrt{s}$. The Weiszfeld algorithm has a fixed point whenever $\boldsymbol{\mu}$ is one of the data points $\boldsymbol{\mu} = \mathbf{x}_k$. As proposed by Vardi and Zhang (2000), one way to deal with this problem is to consider a modified majorizing function $G^*(\boldsymbol{\mu}, \boldsymbol{\mu}_0)$, say, where $G^*(\boldsymbol{\mu}, \boldsymbol{\mu}_0) = G(\boldsymbol{\mu}, \boldsymbol{\mu}_0)$ in (12.18) whenever $\boldsymbol{\mu}_0 \neq \mathbf{x}_k$ for any k and where $G^*(\boldsymbol{\mu}, \boldsymbol{\mu}_0)$ is defined for $\boldsymbol{\mu}_0 = \mathbf{x}_k$ by

$$G^*(\boldsymbol{\mu}, \mathbf{x}_k) = g(\mathbf{x}_k) + w_k|\boldsymbol{\mu} - \mathbf{x}_k| + \sum_{i \neq k} w_i(s_i - s_i^{(0)})\rho'(s_i^0). \tag{12.19}$$

To minimize (12.19) over $\boldsymbol{\mu}$ write $\boldsymbol{\mu} = \mathbf{x}_k + t\mathbf{e}$ where the unit vector $\mathbf{e}$ and the magnitude $t \geq 0$ are to be determined. Then (12.19) can be written in the form

$$G^*(\mathbf{x}_k + t\mathbf{e}, \mathbf{x}_k) = g(\mathbf{x}_k) + w_k t + \frac{1}{2} b_k t^2 + t\mathbf{v}_k^T \mathbf{e}, \tag{12.20}$$

where $b_k = \sum_{i \neq k} w_i / d_{ik}$ and $\mathbf{v}_k$ and d_{ik} are defined in (12.7). First minimize (12.20) over $t \geq 0$ for fixed $\mathbf{e}$ to get $t(\mathbf{e}) = \max(0, -(w_k + \mathbf{v}_k^T \mathbf{e})/b_k)$ yielding

$$G^*(\mathbf{x}_k + t(\mathbf{e})\mathbf{e}) = g(\mathbf{x}_k) - \frac{(w_k + \mathbf{v}_k^T \mathbf{e})^2}{2b_k} \tag{12.21}$$

in (12.20) when $t(\mathbf{e}) > 0$ and $g(\mathbf{x}_k)$ when $t(\mathbf{e}) = 0$. If $\alpha_k > 1$ in (12.8), there is at least one direction $\mathbf{e}$ for which $t(\mathbf{e}) > 0$, and minimizing (12.21) over such directions yields $\mathbf{e} \propto -\mathbf{v}_k$, which by (12.12) is the steepest descent direction for $g(\boldsymbol{\mu})$ from $\mathbf{x}_k$. On the other hand, if $\alpha_k \leq 1$, then we know from Proposition 12.1 that $\mathbf{x}_k$ is already the spatial median and there is no need to update.

In summary, the WOVZ algorithm is given by the following updates:

1. If $\boldsymbol{\mu}$ is not equal to any of the data points, then $\mathbf{M}^*(\boldsymbol{\mu}) = \mathbf{M}(\boldsymbol{\mu})$ is the same as the Weiszfeld update (12.3);
2. If $\boldsymbol{\mu}$ equals a data point $\mathbf{x}_k$ with $\alpha_k > 1$ in (12.8), then the WOVZ update is given by

$$\mathbf{M}^*(\mathbf{x}_k) = \mathbf{x}_k - b_k^{-1}(1 - \alpha_k^{-1})\mathbf{v}_k. \tag{12.22}$$

3. If $\boldsymbol{\mu}$ equals a data point $\mathbf{x}_k$ with $\alpha_k \leq 1$, then $\mathbf{x}_k$ is the spatial median and $\mathbf{M}^*(\mathbf{x}_k) = \mathbf{x}_k$.

Finally we note that the WOVZ update also has an EM-type interpretation when $\boldsymbol{\mu} = \mathbf{x}_k$. In this case we regard all the data $\mathbf{x}_i$, $i \neq k$, as coming from a scale mixture of normals, with missing data $\{Y_i,\ i \neq k\}$, but we regard $\mathbf{x}_k$ as being completely observed. Then the E-step yields (12.19) for minus the log-likelihood.

12.6 One-Dimensional Case

Of course in one dimension there is no need to use an iterative algorithm. It is simpler and more efficient just to find the middle value(s). However, it is still of interest to look at the behavior of the Weiszfeld and WOVZ algorithms. For simplicity we restrict attention to the equally weighted case $w_i \equiv 1$, $1 \leq i \leq n$, with the data points listed in increasing order, $x_1 < x_2 < \cdots < x_n$. An example of the updating function with an odd value of $n = 5$ is given in Fig. 12.1a, and an example with an even value of $n = 6$ is given in Fig. 12.1b. The following features stand out.

1. The updating function is not a monotone function of μ. Instead the plot of $M(\mu)$ vs. μ resembles hanging curtains with kinks at the data points. [Note the distinction between a monotone function and a monotone algorithm!]
2. If n is odd, the median is equal to $x_{(n+1)/2}$. Further, $v_{(n+1)/2} = 0$ in (12.9), so that the linear term disappears. Hence the WOVZ algorithm converges to the median at a *quadratic* rate when μ is close to $x_{(n+1)/2}$. In general EM algorithms converge only at a linear rate, and quadratic convergence is very uncommon.
3. If n is even, the median is not uniquely defined, but can take any value between the two middle data points $x_{n/2}$ and $x_{n/2+1}$. On this middle interval the updating algorithm is fixed, $M(\mu) = \mu$. Further, if μ lies just outside this middle interval, then $M(\mu)$ jumps *inside* the interval whereupon the algorithm stops. Hence this is an unusual example of an EM algorithm which converges in a finite number of steps.

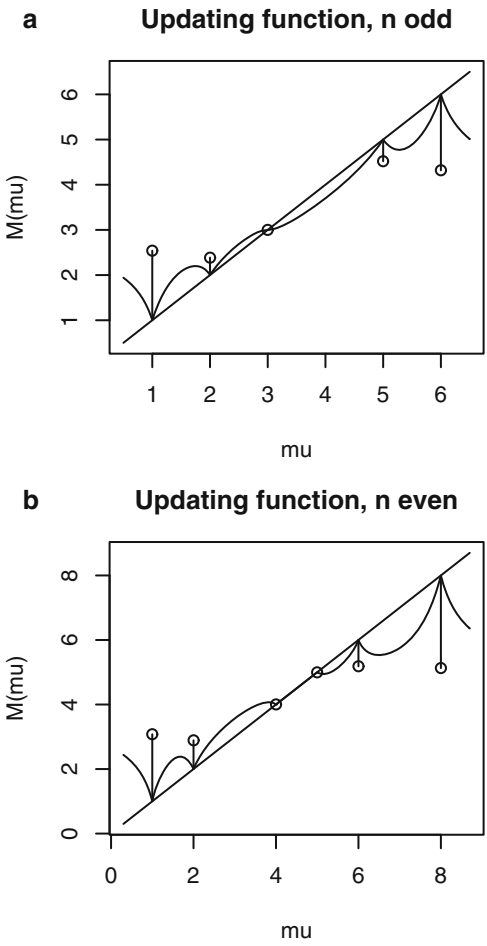

Fig. 12.1 Plots of the updating function $M(\mu)$ in one dimension. The WOVZ updates $M^*(\mu)$ at the data points are indicated by *open circles*. (**a**) The data set consists of $n = 5$ points (1,2,3,5,6). Note that $M(\mu)$ is differentiable and flat at the median, $M'(3) = 0$. (**b**) The data set consists of $n = 6$ points (1,2,4,5,6,8). Note that $M(\mu) = \mu$ between the two middle data points 4 and 5

12.7 Behavior of Newton-Raphson Near a Data Point

For the rest of the paper we restrict attention to dimension $p > 1$. Suppose $\boldsymbol{\mu}$ is close to one of the data points $\mathbf{x}_k$, say and as in Sect. 12.2, write $\boldsymbol{\mu} - \mathbf{x}_k = \epsilon \mathbf{e}_0$, where $\mathbf{e}_0$ is a unit vector and $\epsilon = d_k$ in (12.4). Then the first and second derivatives of g can be written in the form

$$\begin{aligned}
\nabla^T g(\boldsymbol{\mu}) &= w_k \mathbf{e}_0 + \mathbf{v}_k + \mathbf{O}(\epsilon) \\
\nabla\nabla^T g(\boldsymbol{\mu}) &= w_k(\mathbf{I} - \mathbf{e}_0\mathbf{e}_0^T)/\epsilon + \sum_{i \neq k} w_i\{\mathbf{I}/d_{ik} - (\boldsymbol{\mu} - \mathbf{x}_i)(\boldsymbol{\mu} - \mathbf{x}_i)^T/d_{ik}^3\} \\
&= w_k(\mathbf{I} - \mathbf{e}_0\mathbf{e}_0^T)/\epsilon + \mathbf{A}_k + \mathbf{O}(\epsilon),
\end{aligned}$$

that is, the second derivative matrix is $1/\epsilon$ times a rank $p-1$ matrix plus a bounded positive definite matrix

$$\mathbf{A}_k = \sum_{i \neq k} w_i \{\mathbf{I}/d_{ik} - (\mathbf{x}_k - \mathbf{x}_i)(\mathbf{x}_k - \mathbf{x}_i)^T / d_{ik}^3\}, \quad d_{ik} = 1/|\mathbf{x}_k - \mathbf{x}_i|.$$

The $\mathbf{O}(\epsilon)$ terms arise because $\mathbf{v}_k$ and $\mathbf{A}_k$ are evaluated at $\mathbf{x}_k$, not $\boldsymbol{\mu}$.

Using properties of partitioned inverse matrices (e.g., Mardia et al. 1979, p. 459) (to carry out the proof, it is simplest to rotate the coordinate system so that $\mathbf{e}_0 = (1, 0, \ldots, 0)^T$ is a unit vector along the first coordinate axis), it can be checked that

$$\{\nabla \nabla^T g(\boldsymbol{\mu})\}^{-1} = \left(w_k \mathbf{e}_0^T \mathbf{A}_k \mathbf{e}_0\right)^{-1} \mathbf{e}_0 \mathbf{e}_0^T + \mathbf{O}(\epsilon), \tag{12.23}$$

so that the NR update of $\boldsymbol{\mu}$ takes the form

$$\boldsymbol{\mu} - \left(w_k \mathbf{e}_0^T \mathbf{A}_k \mathbf{e}_0\right)^{-1} \mathbf{e}_0 \mathbf{e}_0^T (w_k \mathbf{e}_0 + \mathbf{v}_k) + \mathbf{O}(\epsilon) = \mathbf{x}_k - c\mathbf{e}_0 + \mathbf{O}(\epsilon), \tag{12.24}$$

where $c = \left(\mathbf{e}_0^T \mathbf{A}_k \mathbf{e}_0\right)^{-1} \left(1 + \mathbf{e}_0^T \mathbf{v}_k / w_k\right)$. In other words, the update remains approximately on the line through $\mathbf{x}_k$ in the direction $\pm \mathbf{e}_0$.

In some cases the NR update overshoots. For example, if $\mathbf{e}_0^T \mathbf{v}_k = 0$ and if $g(\mathbf{x}_k + \epsilon \mathbf{e}_0)$ has a minimum in ϵ at $\epsilon = 0$, then the update moves from one side of the line to the other. Further the overshoot remains of order $\mathbf{O}(1)$ as ϵ tends to 0. Of course step-halving can effectively solve the overshoot problem. However, the fact that the NR updates stay on a line is unsatisfactory and can cause NR to converge to a non-optimal data point; see Example 12.1 in Sect. 12.9.

On the other hand, if $\mathbf{e}_0 \propto -\mathbf{v}_k$ lies in the direction of steepest descent, then NR moves away from $\mathbf{x}_k$ by a non-infinitesimal amount and is a sensible update.

Thus we now see a viable way for an optimization algorithm to move away from a non-optimal data point. One step of WOVZ takes $\boldsymbol{\mu}$ to a steepest descent direction from $\mathbf{x}_k$, after which the NR update pulls away from $\mathbf{x}_k$ by a non-infinitesimal amount.

12.8 A Reliable and Efficient Algorithm

The WOVZ algorithm has attractive monotonicity properties ($g(\mathbf{M}(\boldsymbol{\mu})) < g(\boldsymbol{\mu})$ if $\boldsymbol{\mu} \neq \hat{\boldsymbol{\mu}}$). Indeed as Vardi and Zhang (2000) show, it is guaranteed to converge to $\hat{\boldsymbol{\mu}}$. However, the WOVZ algorithm has two major drawbacks.

1. If $\boldsymbol{\mu}$ is near a data point $\mathbf{x}_k$ and α_k is close to 1, then $\mathbf{M}(\boldsymbol{\mu})$ can move very slowly towards (sticky case) or away from (nonsticky case) the data point.

2. Even if $\hat{\boldsymbol{\mu}}$ does not coincide with a data point, the WOVZ algorithm can sometimes converge slowly, a potential problem of all steepest descent and EM algorithms.

To some extent these efficiency problems can be overcome with a line search. The WOVZ update is always in the steepest descent direction. However, the magnitude of the update can occasionally be far too small especially if $\boldsymbol{\mu}$ is close to an (optimal or non-optimal) data point $\mathbf{x}_k$ and if α_k is close to 1. Some examples where a line search would help are given in the next section.

However, even with a line search, steepest descent algorithms are known to be very inefficient if the objective function is quadratic with a highly elliptical rather than spherical shape near its optimum. In this case it is much better to use a second derivative method such as NR. When using NR, some sort of line search is essential to avoid overshooting the optimal update. The choice of line search strategy is not critical, and we shall use a simple step-halving adjustment. That is, define the "update direction" $\mathbf{e}$ (not a unit vector here) by

$$\mathbf{e} = -\left\{\nabla\nabla^T g(\boldsymbol{\mu})\right\}^{-1}\left\{\nabla g(\boldsymbol{\mu})\right\}, \tag{12.25}$$

and consider an update of the form

$$\boldsymbol{\mu} + 2^{-j_1}\mathbf{e}. \tag{12.26}$$

If j_0 is the smallest integer $j \geq 0$ such that $g(\boldsymbol{\mu} - 2^{-j}\mathbf{e}) \leq g(\boldsymbol{\mu})$, then j_1 denotes whichever of j_0 and $j_0 + 1$ yields the smaller value of the objective function.

Unfortunately, NR, even when combined with a line search, should not be used on its own because it is unreliable near data points.

In particular, we saw in Sect. 12.7 that in certain circumstances when $\mathbf{x}_k$ is not the spatial median, the NR updates stay very nearly on the same line through $\mathbf{x}_k$. Thus, even with step-halving, NR can sometimes converge to a non-optimal data point; see Example 12.1 in Sect. 12.9. Another, more minor, problem is that NR is not defined at the data points.

However, by combining approaches, it is possible to fix the bad behavior of both WOVZ and NR near data points. If $\boldsymbol{\mu}$ is close to (or equal to) a data point $\mathbf{x}_k$, then one step of WOVZ takes $\boldsymbol{\mu}$ to a steepest descent direction from $\mathbf{x}_k$, after which the NR update is a sensible update.

We now describe our hybrid algorithm for the spatial median. It turns out not to be important to include a line search in the WOVZ update. Starting with any current estimate $\boldsymbol{\mu}^{(\text{old})}$, let $\boldsymbol{\mu}^{(\text{new})}$ denote the update after a single iteration of the algorithm.

12.8.1 Hybrid Updating Algorithm for the Spatial Median

Step 1: Calculate $d_i = |\boldsymbol{\mu}^{(\text{old})} - \mathbf{x}_i|$, $i = 1, \ldots, n$ and suppose d_k is the smallest value. Consider the possible update given by $\mathbf{x}_k$. If $\alpha_k \leq 1$, then return $\hat{\boldsymbol{\mu}} = \mathbf{x}_k$ and stop.

Step 2: Calculate the WOVZ update (no line search is necessary here).

Step 3: If $\boldsymbol{\mu}^{(\text{old})}$ is distinct from all the data points, also calculate an NR update, adjusted by a step-halving line search.

Step 4: Return as $\boldsymbol{\mu}^{(\text{new})}$ the choice of $\boldsymbol{\mu}$ giving the smallest value of $g(\boldsymbol{\mu})$ from Steps 2 and 3.

Repeat these steps till convergence.

12.9 Examples

Here are several simple examples for the spatial median to illustrate the weaknesses of the WOVZ and NR algorithms on their own (with or without line search). In all cases the hybrid algorithm converges quickly by about 6 iterations. The computing was done in the package `R` (R Core Team 2014) on a machine with an accuracy of about 16 decimal places.

Example 12.1 Consider a dataset of $n = 6$ points in $\mathbb{R}^2$:

$$\begin{array}{cccccc} 100 & 100 & -100 & -100 & 50 & -50 \\ 1 & -1 & 1 & -1 & 0 & 0 \end{array}$$

By symmetry $\hat{\boldsymbol{\mu}} = (0,\ 0)^T$. Note the dataset nearly lies on a straight line, and hence the objective function is nearly flat between the two middle data values $(\pm 50,\ 0)^T$. Further the repulsiveness parameter in (12.8) is $\alpha = 1.04$ for both these two points. Since this value is only just greater than 1, these two data points are only just nonsticky.

- *WOVZ algorithm with no line search.* If we start at $\boldsymbol{\mu} = (49.9,\ 0)^T$, this algorithm stays on the horizontal axis. But it moves away from the data point $(50,\ 0)^T$ very slowly because the corresponding value of α is so close to 1. Indeed, after 200 iterations the algorithm has only reached $(49.893,\ 0)^T$. Further, if started near to $(0,\ 0)^T$, the algorithm still moves extremely slowly because the minimum is so shallow. With a new starting point of $(0.1,\ 0)^T$, the algorithm still only reaches $(0.0990,\ 0)^T$ after 200 iterations. Of course, mathematically the algorithm is guaranteed to converge eventually whatever the starting point.
- *NR with step-halving starting at* $(49.9,\ 0.2)^T$. In this case the algorithm converges to the wrong solution! It converges to the data point $(50,\ 0)^T$. The

problem is that with α so close to 1, there is only a very small window of descent directions from $(50,\ 0)^T$, and NR can have trouble homing in on these directions; see Sect. 12.7. Refining the line search in the NR steps does not help. Since this false convergence is so unexpected with a convex function, we give the first 8 iterations in detail:

Iteration	μ_1	μ_2
0	49.9	0.2
1	49.975613439	0.043951384
2	49.995738930	0.007441284
3	49.998297746	0.002967829
4	49.999578334	0.000734190
5	49.999898708	0.000176305
6	49.999978817	0.000036868
7	49.999998845	0.000002011
8	50.000000096	−0.000000168

- *Hybrid algorithm*. The hybrid algorithm copes easily with both these starting positions. From $(49.9,\ 0)^T$, the algorithm chooses Newton-Raphson updates with step-halving, and converges in just 4 iterations. From $(49.9,\ 0.2)^T$, the hybrid algorithm needs one WOVZ step to get to an approximately steepest descent direction from $(50,\ 0)^T$. After that, NR with step-halving takes over and the algorithm then converges as quickly as before,

Iteration	μ_1	μ_2
0	49.9	0.2
1	49.78	7.2×10^{-7}
2	28.29	6.9×10^{-5}
	⋯	
5	1.4×10^{-6}	1.0×10^{-12}

To set the scene for the next example, we need to use some properties of convex functions. For any point $\mathbf{x}$ distinct from $\hat{\boldsymbol{\mu}}$ and the data points, the objective function $g(\boldsymbol{\mu})$ on the line segment connecting $\mathbf{x}$ to $\hat{\boldsymbol{\mu}}$ must be strictly decreasing. Thus $(\hat{\boldsymbol{\mu}} - \mathbf{x})^T \nabla g(\mathbf{x}) < 0$. In other words $\hat{\boldsymbol{\mu}}^T$ must lie in the half-space $H_{\mathbf{x}} = \{\mathbf{y} \in \mathbb{R}^p : (\mathbf{y} - \mathbf{x})^T \nabla g(\mathbf{x})\} < 0\}$. Similarly, for the non-optimal data points $\mathbf{x}_i$, $\hat{\boldsymbol{\mu}}$ must lie in the half-space $H_i = \{\mathbf{y} \in \mathbb{R}^p :\ (\mathbf{y} - \mathbf{x}_i)^T \mathbf{v}_i < 0\}$, where $\mathbf{v}_i$ was defined in (12.7). In particular, $\hat{\boldsymbol{\mu}}$ must lie in the convex set

$$C = \bigcap H_i \tag{12.27}$$

where the intersection is over indices i corresponding to non-optimal data points.

It is well known that without step-halving, NR can easily diverge if the initial parameter estimate is not close enough to the optimum. Bedall and Zimmermann (1979) tried to avoid the need for step-halving in NR through a two-stage strategy for the spatial median. In the first stage they constructed an simple estimate $\boldsymbol{\mu}^*$ lying in C in (12.27). For the second stage they proposed simple Newton-Raphson updates (i.e. without step-halving). However, as the next example shows, their procedure is not guaranteed to succeed.

Example 12.2 Consider a dataset of $n = 5$ points in $\mathbb{R}^2$:

$$\begin{array}{rrrrr} 10 & 10 & -10 & -85 & 30 \\ 1 & -1 & 0 & 0 & 0 \end{array}$$

In this case it turns out that $\hat{\boldsymbol{\mu}} = (9.42,\ 0)^T$. The output from the first stage of the Bedall–Zimmermann algorithm turns out to be the sample mean $\boldsymbol{\mu}^* = (-9,\ 0)^T$, which lies in C. Unfortunately, using this value as a starting point in a simple Newton-Raphson algorithm does not work.

- *NR without step-halving*. With the starting value $(-9,\ 0)^T$ we get a divergent set of iterations,

Iteration	μ_1	μ_2
0	-9	0
1	3.4×10^4	0
2	-9.96×10^{10}	0

  after which the algorithm fails numerically. If step-halving is included the algorithm converges quickly with no problems, as expected.

The general theory guarantees that the WOVZ algorithm always converges to the spatial median. But in Example 12.1 we showed that without a line search the convergence rate can be very slow. In the next example we show that even with a line search, convergence can still be slow. This drawback is one of the reasons that Newton-Raphson has been included in the hybrid algorithm.

Example 12.3 Consider the dataset with $n = 4$ points in $\mathbb{R}^2$:

$$\begin{array}{rrrr} 100 & -100 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{array}$$

By symmetry $\hat{\boldsymbol{\mu}} = (0,\ 0)^T$. Note that the objective function is highly elliptical near the optimum.

- *WOVZ algorithm with a full line search*. With a starting point of $(1,\ 1)^T$, say, this algorithm behaves as a classic steepest descent algorithm in an elliptical

setting and zig-zags slowly to the optimum. Indeed after 100 iterations it has only reached $(0.001,\ 0.034)^T$.

12.10 Algorithms in R

Algorithms to compute the spatial median have been incorporated in the statistical package `R` (R Core Team 2014). For example, the function `spatial.median` in the library `ICSNP` (Nordhausen et al. 2012) uses the WOVZ algorithm and is recommended in Oja (2010, p. 71).

The library `pcaPP` (Filzmoser et al. 2013) contains functions based on several algorithms. These include `l1median_NLM`, which is based on the standard `R` optimization routine `nlm`; `l1median_VaZh`, which uses the WOVZ algorithm; and `l1median_HoCr`, which uses the WOVZ algorithm with step halving (Hössjer and Croux 1995).

The following simple examples illustrate some issues that can arise with these algorithms. In all cases the standard options have been used.

- The severe eccentricity of the objective function in Example 12.1 of Sect. 12.9 means that an algorithm can sometimes stop too soon. With an initial value $\mathbf{x} = (49.9,\ 0)^T$, `l1median_NLM` computes the spatial median as $\hat{\boldsymbol{\mu}} = (0.19,\ 0)^T$ instead of $(0,\ 0)^T$, after 5 iterations. This is a fairly minor issue since the objective function is nearly constant on a long valley.
- Again in Example 12.1 of Sect. 12.9, the fact that for the data point $(50,\ 0)^T$ the repulsiveness parameter $\alpha = 1.04$ is so close to 1 means that the WOVZ-based algorithms move very slowly when started near the data point. With a starting point $(49.9,\ 0)^T$, the functions `l1median_VaZh` and `l1median_HoCr` behave as the WOVZ algorithm in Example 12.1, stopping after the default maximum number of iterations, 200, without having converged, with the same poor estimate as found there, $\hat{\boldsymbol{\mu}} = (49.893,\ 0)^T$. The function `spatial.median` explicitly notes that it has not converged and fails to produce an estimate. However, when started even closer to the data point $(50, 0)^T$, these functions can falsely declare convergence before they get started. The functions `l1median_VaZh` and `spatial.median`, starting at $(49.999,\ 0)^T$, stop after one iteration and falsely declare $\hat{\boldsymbol{\mu}}$ to be equal to the starting point. The function `l1median_HoCr` does the same if started from the closer value $(49.99999,\ 0)^T$.
- Being an all-purpose minimization algorithm, `l1median_NLM` does not take full advantage of the structure of the problem. For example, consider the dataset

$$\begin{array}{cccccc} 1 & -1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \end{array}$$

This is a delicate dataset. The spatial median lies at the sixth data point, which is the origin $(0,\ 0)^T$. However, the repulsiveness parameter of the origin is $\alpha = 1$, which lies on the sticky/nonsticky boundary. For this example `l1median_NLM` can be very slow. Starting from $(1,\ 1)^T$ `l1median_NLM` takes 120 iterations to converge to the right answer, whereas the hybrid algorithm converges in just 2 iterations.

Fritz et al. (2012) gave a numerical comparison of various algorithms, focusing on speed and accuracy. In general, it was found that the function `l1median_NLM` came out best. However, although `l1median_NLM` performs well in general, the toy examples here demonstrate how there can be problems for certain datasets with certain starting points.

12.11 Convex Optimization

Since g is a convex function, it is also possible to draw on general methods of convex optimization. Here we sketch some possible approaches, though we will not investigate these methods in detail in this paper.

In one approach, a non-smooth optimization problem is replaced by a smooth optimization problem, depending on a tuning parameter η, say, for which standard (e.g., Newton-type) methods can be applied. As $\eta \to 0$ the solution to the smoothed problem tends to the solution of the original problem. The simplest version of this idea is to use the perturbed objective function

$$g_\eta(\boldsymbol{\mu}) = \sum_{i=1}^{n} \{|\mathbf{x}_i - \boldsymbol{\mu}|^2 + \eta^2\}^{1/2}.$$

Provided $\eta > 0$, this function is twice-differentiable. Kärkkäinen and Äyrämö (2005) investigate a modified Weiszfeld algorithm with acceleration.

Another approach, too involved to describe here, is to adapt primal–dual interior point methods from linear programming problems. See, e.g., Lobo et al. (1998), Andersen (1996), Andersen et al. (2000) for details. Other approaches include alternating direction method of multipliers (ADMM) methods such as FISTA (Beck and Teboulle 2009) and the Chambolle and Pock (2011) method.

However, the level of complexity in these approaches seems unnecessary for the simple problem of finding the spatial median. Hence, for this paper we have limited attention to simpler algorithms. On the other hand, a major advantage of more general convex optimization methods is that they can be applied to a much wider class of optimization problems (see, e.g., Lobo et al. 1998). In particular, Valkonen and Kärkkäinen (2010) discuss problems related to the spatial median when there are missing coordinates for some of the data.

12.12 Conclusions

Two standard algorithms in the statistical literature used to find the spatial median are WOVZ and NR. Both algorithms can have difficulties at and near the data points. These difficulties have not always been fully appreciated in the literature.

The problems with WOVZ and NR can be fixed by using a hybrid algorithm that takes the better choice at each update. The good properties of the hybrid algorithm have been demonstrated by mathematical expansions and simple numerical examples.

References

Andersen, K.D.: An efficient Newton barrier method for minimizing a sum of Euclidean norms. SIAM J. Optim. **6**, 74–95 (1996)

Andersen, K.D., Christiansen, E., Conn, A.R., Overton, M.L.: An efficient primal-dual interior-point method for minimizing a sum of Euclidean norms. SIAM J. Sci. Comput. **22**, 243–262 (2000)

Andrews, D.F., Mallows, C.L.: Scale mixture of normal distributions. J. R. Stat. Soc. Ser. B **36**, 99–102 (1974)

Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. **2**, 183–202 (2009)

Bedall, F.K., Zimmermann, H.: The mediancentre. Appl. Stat. **28**, 325–328 (1979)

Brown, B.M.: Statistical uses of the spatial median. J. R. Stat. Soc. Ser. B **45**, 25–30 (1983)

Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vision **40**, 120–145 (2011)

Dempster, A.P., Laird, N.M., Rubin, D,B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B **39**, 1–38 (1977)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Iteratively reweighted least squares for linear regression when errors are Normal/Independent distributed. In: Krishnaiah, P.R. (ed.) Multivariate Analysis-V, pp. 35–57. North-Holland, Amsterdam (1980)

Feller, W.: An Introduction to Probability Theory and its Applications, vol. II. Wiley, New York (1966)

Filzmoser, P., Fritz, H., Kalcher, K.: pcaPP: Robust PCA by Projection Pursuit. R package version 1.9–49 (2013). http://CRAN.R-project.org/package=pcaPP

Fritz, H., Filzmoser, P., Croux, C.: A comparison of algorithms for the multivariate L_1 median. Comput. Stat. **27**, 393–410 (2012)

Gower, J.C.: The mediancentre. Appl. Stat. **23**, 466–470 (1974)

Haldane, J.B.S.: Note on the median of a multivariate distribution. Biometrika **35**, 414–415 (1948)

Hössjer, O., Croux, C.: Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter. Nonparametric Stat. **4**, 293–308 (1995)

Huber, P.J.: Robust Statistics. Wiley, Chichester (1981)

Hunter, D.R., Lange, K.: A tutorial on MM algorithms. Am. Stat. **58**, 30–37 (2004)

Kärkkäinen, T., Äyrämö, S. On computation of spatial median for robust data mining. In: Schilling, R., Hasse, W., Periaux, J., Baier, H., Bugeda, G. (eds.) Evolutionary and Deterministic Methods for Design, Optimization and Control with Applications to Industrial and Societal Problems, EUROGEN 2005. FLM, Munich (2005)

Kuhn, H.W.: A note on Fermat's problem. Math. Program. **4**, 98–107 (1973)

Lange, K., Hunter, D.R., Yang, I.: Optimization transfer using surrogate objective functions (with discussion). J. Comput. Graph. Stat. **9**, 1–59 (2000)
Lange, K., Sinsheimer, J.S.: Normal/Independent distributions and their applications in robust regression. J. Comput. Graph. Stat. **2**, 175–198 (1993)
Lobo, M.S., Vandenberghe, L., Boyd, S., Lebret, H.: Applications of second-order cone programming. Linear Algebra Appl. **284**, 193–228 (1998)
Mardia, K.V., Kent, J.T., Bibby, J.M.: Multivariate Analysis. Academic, London (1979)
McLachlan, G., Krishnan, T.: The EM Algorithm and Extensions, 2nd edn. Wiley, Hoboken (2008)
Nordhausen, K., Sirkiä, S., Oja, H., Tyler, D.E.: ICSNP: Tools for Multivariate Nonparametrics. R package version 1.0-9 (2012). http://CRAN.R-project.org/package=ICSNP
Oja, H.: Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks. Springer, New York (2010)
Ostresh, L.M.: On the convergence of a class of iterative methods for solving the Weber location problem. Oper. Res. **26**, 597–609
R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria (2014)
Small, C.G.: A survey of multidimensional medians. Int. Stat. Rev. **58**, 263–277 (1990)
Valkonen, T., Kärkkäinen, T.: Clustering and the perturbed spatial median. Math. Comput. Model. **52**, 87–106 (2010)
Vardi, Y., Zhang, C.-H.: The multivariate L_1-median and associated data depth. Proc. Natl. Acad. Sci. **97**, 1423–1426 (2000)
Wu, C.F.J.: On the convergence properties of the EM algorithm. Ann. Stat. **11**, 95–103 (1983)

Chapter 13
L_1-Regression for Multivariate Clustered Data

Jaakko Nevalainen and Denis Larocque

Abstract In this chapter, we are considering L_1-type estimation for multivariate clustered data. Although valid, using the direct L_1 estimation of the regression coefficients in the clustered data setting is likely to lack efficiency since it does not use the intracluster correlation structure. A transformation–retransformation method is proposed to overcome this problem. This method first transforms the original model in an attempt to eliminate the intracluster correlation. Secondly, the L_1 estimates are obtained with the transformed data, which are then transformed back to the original scale. One particular implementation of this method is investigated in a simulation study which shows that it is more efficient than using the direct L_1 estimators.

Keywords Clustered data • L_1-Estimation • Multivariate analysis • Nonparametric statistics • Regression • Spatial sign

13.1 Introduction

In this chapter, we consider L_1-regression models for clustered data with a multivariate response. For a univariate response, Jung and Ying (2003) proposed a generalization of the Wilcoxon–Mann–Whitney statistic for analyzing repeated measurements data. The estimating function is based on the unweighed ranks of the residuals, which is equivalent to the method proposed by Jurečková (1969, 1971) for independent observations. In order to recover some of the information present in the clustering structure, Wang and Zhu (2006) generalized this approach by partitioning the ranks into between- and within-subject ranks. Two estimators are obtained and then combined in an optimal way. However, to get the combined estimator, an

J. Nevalainen (✉)
School of Health Sciences, University of Tampere, 33014 Tampere, Finland
e-mail: jaakko.nevalainen@uta.fi

D. Larocque
Department of Decision Sciences, HEC Montréal, 3000 chemin de la Côte-Sainte-Catherine, Montréal, QC, Canada H3T 2B1
e-mail: denis.larocque@hec.ca

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_13

estimation of the covariance matrix of the two estimating functions is required. To achieve this, a resampling method is used in Wang and Zhu (2006). Fu et al. (2010) proposed a smoothing method to avoid this computationally intensive approach. In another attempt to use the clustering structure, Wang and Zhao (2008) proposed a weighted version of the loss function, where the weights are functions of the cluster sizes. Their approach is related to the one proposed in Datta and Satten (2005) for the two-sample problem. Fu and Wang (2012) argue that the Wang and Zhao (2008) approach performs well for cluster-level covariates but not necessarily for within-cluster covariates. They derive a new optimal rank-based estimating functions in terms of asymptotic variance of regression parameter estimators. Finally, Kloke et al. (2009) study R-estimators of the fixed effects in an experiment done over clusters, blocks, groups, or subjects, including for example, repeated measure designs, split plot designs, multicenter clinical trials, randomized block designs, and two-stage cluster samples.

All the articles above are aimed at the univariate response case. Nonparametric methods for multivariate data, and especially, methods based on spatial signs and ranks have been developed extensively in the last 20 years (Oja 2010). They are also available for the user through the R package `MNM` (Nordhausen and Oja 2011). Moreover, specialized methods for multivariate responses and clustered data have also been developed; see Nevalainen et al. (2010) and the references therein. However, for clustered data, the available methods are limited so far to the one, two, and several samples cases. In this chapter we propose an L_1-type (spatial sign) estimation method for a regression setting with multivariate clustered data and investigate it in a simulation study.

13.2 A Multivariate Multiple Linear Regression Model for Clustered Data

Let $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^\top$ be a sample of p-variate ($p > 1$) random response vectors with sample size n. The data are assumed to be clustered with a total of d clusters. The cluster memberships are given by the $n \times d$ matrix $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\top$:

$$(\mathbf{Z})_{ij} = \begin{cases} 1, \text{ if the } i\text{th observation is from cluster } j; \\ 0, \text{ otherwise.} \end{cases}$$

It is useful to note that

$$\left(\mathbf{Z}\mathbf{Z}^\top\right)_{ij} = \begin{cases} 1, \text{ if the } i\text{th and the } j\text{th observation are from the same cluster;} \\ 0, \text{ otherwise,} \end{cases}$$

and that $\mathbf{Z}^\top\mathbf{Z}$ is a $d \times d$ diagonal matrix with the cluster sizes on the diagonal, say, $m_1, \ldots, m_d$. We also write $\mathbf{1}_n$ for a column n-vector of ones, vec($\mathbf{Y}$) for the vector obtained by stacking the columns of $\mathbf{Y}$, and $\otimes$ for the Kronecker product.

Consider the multivariate multiple linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E},$$

where

- $\mathbf{X}$ is an $n \times q$ design matrix for explanatory variables with the first column consisting of 1's;
- $\boldsymbol{\beta}$ is the $q \times p$ matrix of regression coefficients;
- $\mathbf{E} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)^\top$ is an $n \times p$ matrix of random errors,

stating that the responses are linearly related to the explanatory variables, and $\mathbf{E}$ is a matrix of random errors with

$$\mathbf{Cov}\left(\mathrm{vec}(\mathbf{E}^\top)\right) = \boldsymbol{\Omega} \otimes \boldsymbol{\Sigma}.$$

Here $\boldsymbol{\Sigma} = E(\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^\top)$ and $\boldsymbol{\Omega} = \{\rho_{ij}\}$ is a matrix consisting of intracluster correlations, with unit entries on the diagonal. We thus assume that $E(\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_j^\top) = \rho_{ij}\boldsymbol{\Sigma}$, where $\rho_{ij} \neq 0$ if $(\mathbf{Z}\mathbf{Z}^\top)_{ij} = 1$, and $\rho_{ij} = 0$, otherwise.

Typically, the regression coefficients of the model are estimated by ordinary least squares based on a minimization of an L_2 criterion function, or by maximum likelihood relying on a multivariate normality assumption on the random errors. These two solutions are in general not the same with clustered data. Under heavy-tailed error distributions these estimates are inefficient, and may be vulnerable to outliers. In such circumstances, a fit based on an L_1-criterion function may be preferable.

13.3 Estimation Based on an L_1-Criterion Function

The goal is to estimate the unknown $\boldsymbol{\beta}$ matrix of regression coefficients by minimizing an L_1 norm

$$D_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(|\mathbf{y}_i - \boldsymbol{\beta}^\top \mathbf{x}_i| - |\mathbf{y}_i| \right).$$

This leads to a multivariate *least absolute deviation (LAD) estimate* of $\boldsymbol{\beta}$. If the residuals lie in a genuinely p-dimensional space, the resulting estimate $\hat{\boldsymbol{\beta}}$ solves the estimating equation

$$\mathbf{U}(\hat{\boldsymbol{\beta}})^\top \mathbf{X} = \mathbf{0},$$

where

$$\mathbf{U}_i(\boldsymbol{\beta}) = (\mathbf{y}_i - \boldsymbol{\beta}^\top \mathbf{x}_i)/|\mathbf{y}_i - \boldsymbol{\beta}^\top \mathbf{x}_i|,\ i = 1, \ldots, n$$

is the spatial sign of the residual at $\boldsymbol{\beta}$ and $\mathbf{U}(\boldsymbol{\beta}) = (\mathbf{U}_1(\boldsymbol{\beta}), \ldots, \mathbf{U}_n(\boldsymbol{\beta}))^\top$ is the corresponding matrix of residual spatial signs (Oja 2010). These L_1 estimates of regression coefficients are quite natural and not difficult to compute (Nordhausen and Oja 2011).

13.4 Transformation–Retransformation L_1 Regression Estimates

It is possible to use the L_1 norm to directly estimate the parameters also in the clustered data case. Compared to a setting with i.i.d. random errors, the limiting distribution would only require a correction in the variance terms. This permits a valid analysis. However, the estimate as such suffers from one important shortcoming: it makes no use of the underlying and known cluster structure. A reasonable concern is that it may be an inefficient approach; recall that the (optimal) maximum likelihood estimates for linear models in the univariate normal case use the covariance structure as an essential ingredient.

Suppose first that the covariance matrix was known and again has the general form

$$\mathbf{Cov}(\mathrm{vec}(\mathbf{E}^\top)) = \boldsymbol{\Omega} \otimes \boldsymbol{\Sigma}.$$

For example, the "compound symmetry" covariance structure,

$$\mathbf{Cov}(\mathrm{vec}(\mathbf{E}^\top)) = \mathbf{I}_n \otimes \boldsymbol{\Sigma} + (\mathbf{Z}\mathbf{Z}^\top - \mathbf{I}_n) \otimes \rho\boldsymbol{\Sigma} = (\mathbf{I}_n + \rho(\mathbf{Z}\mathbf{Z}^\top - \mathbf{I}_n)) \otimes \boldsymbol{\Sigma}$$

falls into this class of structures.

Given a pre-specified covariance structure $\boldsymbol{\Omega} \otimes \boldsymbol{\Sigma}$, the original estimation problem can be rewritten as

$$\begin{aligned}
\mathbf{Y} &\to \mathbf{Y}_0 = \boldsymbol{\Omega}^{-1/2}\mathbf{Y}\boldsymbol{\Sigma}^{-1/2} \\
\mathbf{X} &\to \mathbf{X}_0 = \boldsymbol{\Omega}^{-1/2}\mathbf{X} \\
\boldsymbol{\beta} &\to \boldsymbol{\beta}_0 = \boldsymbol{\beta}\boldsymbol{\Sigma}^{-1/2} \\
\mathbf{E} &\to \mathbf{E}_0 = \boldsymbol{\Omega}^{-1/2}\mathbf{E}\boldsymbol{\Sigma}^{-1/2}
\end{aligned}$$

This postulates a new regression model $\mathbf{Y}_0 = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{E}_0$ in which, if the covariance structure is correctly specified, the random errors are uncorrelated. Multiplication from the left attempts to eliminate the intracluster correlation, and multiplication from the right is aimed to standardize the marginal distributions. For the transformed data set on the right-hand side, it is reasonable to conduct ordinary L_1 estimation of regression coefficients as before. Therefore, the estimating equation is

$$\mathbf{U}_0\left(\boldsymbol{\beta}\boldsymbol{\Sigma}^{-1/2}\right)^\top \boldsymbol{\Omega}^{-1/2}\mathbf{X} = \mathbf{0},$$

where $\mathbf{U}_0$ now consists of the spatial signs of the residuals on the transformed scale, i.e., the rows of $\boldsymbol{\Omega}^{-1/2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\boldsymbol{\Sigma}^{-1/2}$. As a final step, the estimates of the regression parameters in the original model are obtained by back-transformation $\hat{\boldsymbol{\beta}}_{trt} = \hat{\boldsymbol{\beta}}\boldsymbol{\Sigma}^{1/2}$.

This procedure has potential for improved efficiency. Similar idea of a working, user-specified correlation structure is met in the framework of generalized estimating equations (Diggle et al. 2002). In that setup, the estimates of the regression coefficients are consistent even if the working correlation structure is misspecified. Simulations will be used to investigate the merits of this method in the next section.

13.5 Simulation Study

We next investigate the finite sample efficiency of the proposed method by simulation studies involving four competing approaches to the problem.

13.5.1 Practical Implementation of the Transformation–Retransformation L_1 Method

We are assuming that the working covariance structure is $E(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_j^\top) = \rho_{ij}\boldsymbol{\Sigma}$, where $\rho_{ij} = \rho$ if $(\mathbf{Z}\mathbf{Z}^\top)_{ij} = 1$ and $\rho_{ij} = 0$ otherwise. We thus assume a compound symmetry structure with the same correlation (ρ) for each response. In this respect the compound symmetry structure is a special case; other covariance structures would not in general have this property. The robustness of the method when this assumption is not met is investigated in the simulations of Sect. 13.5.4.

A particular implementation of the transformation–retransformation L_1 method for this setting is as follows. To estimate ρ, we take the average of the intracluster correlation estimate obtained from separate linear mixed models to each response. More precisely, let $\hat{\rho}_i$ be the estimated value of the intracluster correlation obtained from fitting a linear mixed model to the ith response with a random intercept at the cluster level, and using all the covariates. Then $\hat{\rho} = (1/p)\sum_{i=1}^{p}\hat{\rho}_i$.

To estimate $\boldsymbol{\Sigma}$, we use the sample covariance matrix of the residuals obtained from the same response-wise linear mixed models.

13.5.2 Design of the Simulation

Data are generated according to the following multivariate linear mixed model with $p = 3$ or 7 responses and d clusters:

$$\mathbf{Y}_{ij} = X_{1ij}\mathbf{1}_p + X_{2ij}\mathbf{1}_p + X_{3ij}\mathbf{1}_p + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij}, \quad i = 1, \ldots, d, \quad j = 1, \ldots, m_i.$$

In this model, $\mathbf{Y}_{ij}$ is the p-variate response vector for the jth observation in cluster i, the $\boldsymbol{\alpha}_i$'s are i.i.d. p-variate cluster effects (random intercept), and $\boldsymbol{\epsilon}_{ij}$'s are i.i.d. p-variate individual error terms. The three covariates (the X's) are i.i.d. from $N(0, 1)$. The covariates, random intercepts, and error terms are all independent. Hence the true matrices of regression coefficients are

$$\boldsymbol{\beta} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

for the $p = 3$ and 7 cases, respectively.

The cluster design consists in 5 clusters of size 2, 5 clusters of size 3, 5 clusters of size 4, 5 clusters of size 5 and 5 clusters of size 6, for a total of 25 clusters and 100 observations.

The $\boldsymbol{\alpha}_i$'s and $\boldsymbol{\epsilon}_{ij}$'s are generated from either the normal or the t_3 distribution with mean vector $\mathbf{0}$. In all cases, the scale matrix of these distribution has the form $\rho\mathbf{I}_p$ for $\boldsymbol{\alpha}_i$ and $(1-\rho)\mathbf{I}_p$ for $\boldsymbol{\epsilon}_{ij}$. We then let ρ vary between 0 and 0.95 by steps of 0.05.

Four estimation methods are compared.

1. Transformation–retransformation L_1 method assuming the compound symmetry structure with equal ρs. The estimation of the parameters is described in Sect. 13.5.1. This is the proposed method.
2. Random intercept linear mixed models fitted separately to each response.
3. Basic L_1 regression applied directly to the data, neglecting the intracluster correlation.
4. Transformation–retransformation L_1 method assuming the compound symmetry structure with equal ρs. But this time we use the true values of ρ. However, we still estimate $\boldsymbol{\Sigma}$ as explained in Sect. 13.5.1. This method is not feasible in practice because ρ will likely never be known. But we use it as a benchmark to investigate the effect of having to estimate ρ.

The number of simulation runs is 500 for each configuration. All computations are performed in `R` (R Core Team 2013). The linear mixed models are fitted with the `lme` function in the `nlme` package (Pinheiro et al. 2014). The L_1 regressions are performed with the `mv.l1lm` function in the `MNM` package (Nordhausen and Oja 2011).

13.5.3 Results

For each estimation method r $(= 1, 2, 3, 4)$, the performance criterion is

$$P_r = (1/500) \sum_{i=1}^{500} \text{vec}(\hat{\boldsymbol{\beta}}_{ri} - \boldsymbol{\beta})' \text{vec}(\hat{\boldsymbol{\beta}}_{ri} - \boldsymbol{\beta}),$$

where $\hat{\boldsymbol{\beta}}_{ri}$ is the estimation from this method for the ith simulation run. The results are summarized in Fig. 13.1. In each plot, the Y-axis gives the efficiency of each method relative to the proposed method (method 1 in Sect. 13.5.2), as a function of ρ. More precisely, it is P_1/P_r for $r = 2, 3, 4$. Hence, the proposed method is more efficient than the other one when the relative efficiency is below 1 and less efficient when it is above 1. The upper-left plot corresponds to the three-variate case where both the random intercepts and the errors are normally distributed. As expected, the linear mixed model (method 2) is a bit more efficient than the proposed method in this case, and the efficiency remains constant over the range of ρ. We can also see clearly the effect of neglecting the intracluster correlation. The performance of the basic L_1 regression (method 3) dramatically worsens as ρ increases. We also see that using the true value of ρ (method 4) has the same performance as the proposed method. Hence, nothing is lost by having to estimate this parameter.

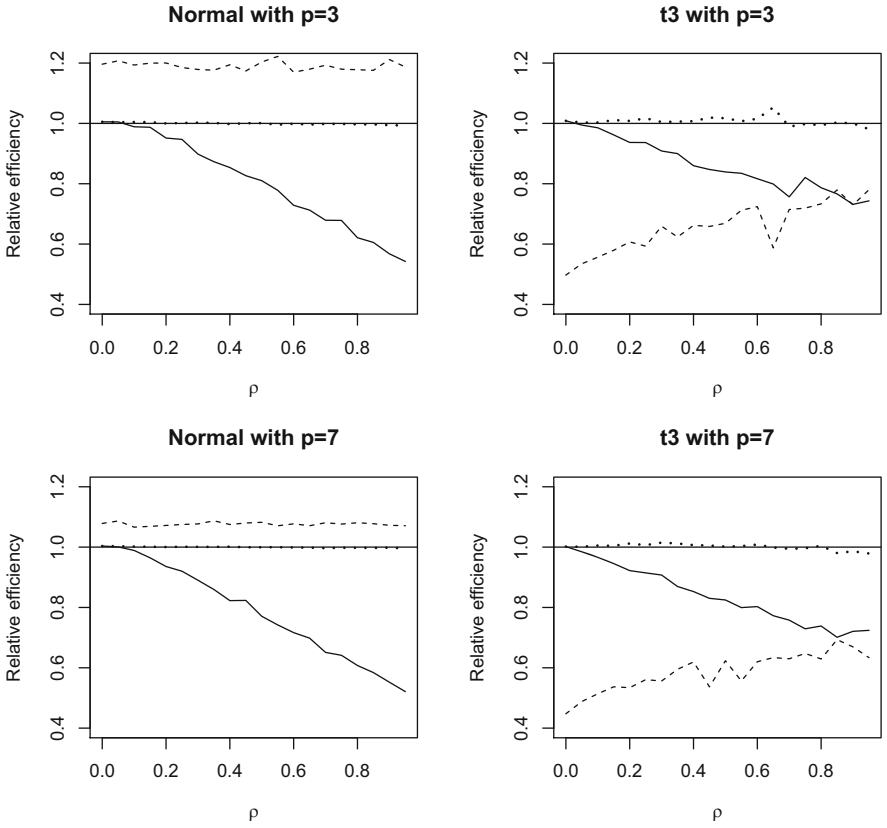

Fig. 13.1 Efficiency relative to the transformation–retransformation L_1 method with estimated ρ. Equal ρ case. Basic L_1 is the *full line* (–), mixed model is the *dashed line* (- - -), transformed L_1 with true ρ is the *dotted line* (⋯)

The upper-right plot corresponds to the same situation but with random intercepts and errors distributed as t_3. This time, the proposed method is more efficient than the linear mixed model. Again, this was expected since L_1 based methods are more efficient for heavier-tailed distributions. However, the difference between the two is more pronounced for smaller values of ρ. The performance of the basic L_1 regression worsens as ρ increases and having to estimate ρ does not hurt the performance.

The lower plots present the corresponding results for the seven-variate cases. For the normal case (lower-left), the same patterns as for the three-variate case occur, except that the difference between the proposed method and the linear mixed model is even smaller. For the t_3 case, (lower-right), the situation is similar and the same patterns as for the three-variate case also occur. Hence, from these limited results, the proposed transformation–retransformation L_1 method is clearly preferable to the direct L_1 method.

13.5.4 Additional Simulations Under Unequal ρs Scenarios

The results presented previously showed that the transformation–retransformation L_1 method is more efficient than the direct L_1 method. However, the particular implementation of the method used in the simulation assumes a compound symmetry structure with the same correlation for each response. Hence, it is natural to ask if it still performs well when this is not true. To investigate this, we used the same scenarios as before, except that we allowed the intracluster correlations to vary for each response. More precisely, the $\boldsymbol{\alpha}_i$'s and $\boldsymbol{\epsilon}_{ij}$'s are still generated from either the normal or the t_3 distribution with mean vector $\mathbf{0}$. But this time, the scale matrix of these distribution has the form $\text{diag}(\rho_1, \ldots, \rho_p)$ for $\boldsymbol{\alpha}_i$ and $\text{diag}((1-\rho_1), \ldots, (1-\rho_p))$ for $\boldsymbol{\epsilon}_{ij}$. When $p = 3$, we fix $\rho_2 = 0.5$, $\rho_1 = \rho$, $\rho_3 = 1-\rho$, and we let ρ vary between 0.05 and 0.5 by steps of 0.05. When $p = 7$, we fix $\rho_4 = 0.5$, $\rho_1 = \rho_2 = \rho_3 = \rho$, $\rho_5 = \rho_6 = \rho_7 = 1 - \rho$, and we let ρ vary between 0.05 and 0.5 by steps of 0.05. Hence, the greatest variance among the ρs occur when $\rho = 0$. When $\rho = 0.5$, we fall back to the equal ρs case. Note that method 4 is not applicable with these scenarios. Therefore, only the first three are compared.

Figure 13.2 presents the results. The upper-left plot corresponds to the three-variate case where both the random intercepts and the errors are normally distributed. We see that the proposed method continues to be more efficient than the direct L_1 method even though its working correlation structure (compound symmetry with equal ρs) is not true. The comparison with the linear mixed model is similar to what we have in Fig. 13.1. We can only see a little increase in the curve when $\rho = 0.05$, which corresponds to the case where the variance among the ρs is the greatest, and thus where we are the farthest away from the working correlation. Hence it seems that the proposed method is quite robust to this type of departure from the assumptions.

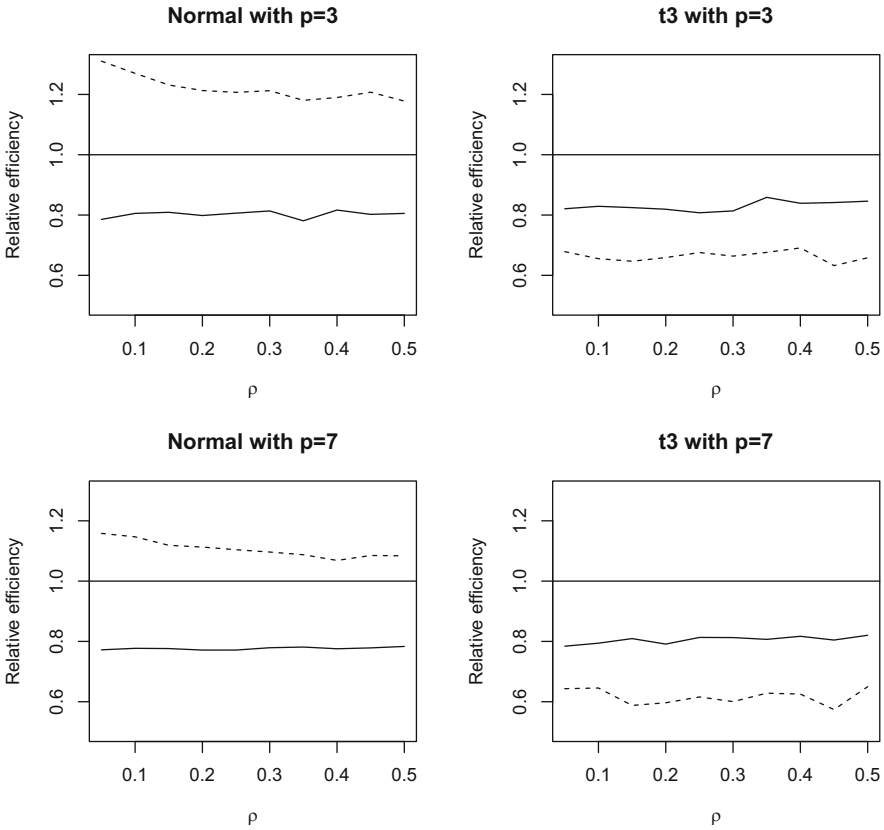

Fig. 13.2 Efficiency relative to the transformation–retransformation L_1 method with estimated ρ. Unequal ρ case. Basic L_1 is the *full line* (–), mixed model is the *dashed line* (- - -)

The upper-right plot corresponds to the same situation but with random intercepts and errors distributed as t_3. This time, the proposed method is always more efficient than both the direct L_1 method and the linear mixed model. Similar findings can be obtained from the lower plots, which present the corresponding results for the seven-variate cases.

The conclusion from this limited simulation study is that the transformation–retransformation L_1 method is very promising as it seems more efficient than the direct L_1 method. Moreover, the advantage over the direct L_1 method seems to hold even when the working correlation structure is not well specified. Hence, the partial information recovered by the transformation–retransformation L_1, even if coming from a wrongly specified working correlation structure, is still useful enough to beat the direct L_1 method.

13.6 Concluding Remarks

The goal of this paper was to show the potential of a novel L_1 norm estimation method of regression coefficients for clustered data. Unlike using the L_1 norm directly on the data, the proposed transformation–retransformation method uses the clustering structure to produce more efficient estimates, as shown in a simulation study. Hence this method deserves to be investigated further. The next logical step is to study the theoretical properties of the proposed method. More precisely, we are planning to derive its asymptotic properties, including calculations of asymptotic efficiencies. Moreover, we used a simple estimation of ρ, based on response-wise linear mixed models in the particular implementation used in the simulation study. However, using an L_1-type method would be more natural and we are also planning to develop such a method.

Acknowledgements This research was supported by NSERC, FRQNT, and the Academy of Finland. The authors also thank two anonymous reviewers for helpful comments.

References

Datta, S., Satten, G.A.: Rank-sum tests for clustered data. J. Am. Stat. Assoc. **471**, 908–915 (2005)
Diggle, P., Heagerty, P., Liang, K.-Y., Zeger, S.: Analysis of Longitudinal Data, 2nd edn. Oxford University Press, Oxford (2002)
Fu, L., Wang, Y.-G.: Efficient estimation for rank-based regression with clustered data. Biometrics **68**, 1074–1082 (2012)
Fu, L., Wang, Y.-G., Bai, Z.: Rank regression for analysis of clustered data: a natural induced smothing approach. Comput. Stat. Data Anal. **54**, 1036–1050 (2010)
Jung, S.-H., Ying, Z.: Rank-based regression with repeated measurements data. Biometrika **90**, 732–740 (2003)
Jurečková, J.: Asymptotic linearity of a rank statistic in regression parameter. Ann. Math. Stat. **40**, 1889–1900 (1969)
Jurečková, J.: Non-parametric estimate of regression coefficients. Ann. Math. Stat. **42**, 1328–1338 (1971)
Kloke, J.D., McKean, J.W., Mushfiqur Rashid, M.: Rank-based estimation and associated inferences for linear models with cluster correlated errors. J. Am. Stat. Assoc. **485**, 384–390 (2009)
Nevalainen, J., Larocque, D., Oja, H., Pörsti, I.: Nonparametric analysis of clustered multivariate data. J. Am. Stat. Assoc. **490**, 864–872 (2010)
Nordhausen, K., Oja, H.: Multivariate L_1 methods: the package MNM. J. Stat. Softw. **43**, 1–28 (2011)
Oja, H.: Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks. Springer, New York (2010)
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team: nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-117. http://CRAN.R-project.org/package=nlme (2014)
R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. http://www.R-project.org (2013)
Wang, Y.-G., Zhao, Y.: Weighted rank regression for clustered data analysis. Biometrics **64**, 39–45 (2008)
Wang, Y.-G., Zhu, M.: Rank-based regression for analysis of repeated measures. Biometrika **93**, 459–464 (2006)

Chapter 14
Robust Variable Selection and Coefficient Estimation in Multivariate Multiple Regression Using LAD-Lasso

Jyrki Möttönen and Mikko J. Sillanpää

Abstract Univariate and multivariate lasso estimation methods are highly sensitive to outlying observations because of the sum of squared norms term in the objective function. Using sum of norms (least absolute deviations, LAD) instead of sum of squared norms gives us a considerably more robust estimate for the regression coefficients. In this paper we combine LAD with the multivariate lasso method and illustrate its estimation using simulated data set that are similar to those typically seen in association genetics. We will shortly consider also how the significance testing is done for non-zero coefficients and how the tuning parameter value can be determined.

Keywords Association genetics • Multivariate multiple regression • LASSO • Least absolute deviation

14.1 Introduction

In many applications of genomics, only a small subset of explanatory variables is likely to be important a posteriori with respect to the outcome variable. However, the number of explanatory variables needed to be considered a priori in the regression model is much higher than the number of individuals in the sample. This necessitates the use of regularization in the estimation or prior information in the model to obtain reasonable estimates of the regression coefficients for important explanatory variables.

J. Möttönen (✉)
Department of Social Research, University of Helsinki, P.O. Box 18, 00014 Helsinki, Finland
e-mail: jyrki.mottonen@helsinki.fi

M.J. Sillanpää
Department of Mathematical Sciences and Biocenter Oulu, University of Oulu, P.O. Box 3000, 90014 Oulu, Finland
e-mail: mikko.sillanpaa@oulu.fi

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_14

The regularization can be done either by using shrinkage estimation (Tibshirani 1996) or by using simultaneous estimation and variable selection (O'Hara and Sillanpää 2009). Shrinkage estimation operates by automatically shrinking spurious (un-important) coefficients towards zero while coefficients of important covariates obtain non-zero values. Recently proposed regularization methods for linear regression include, for example, lasso (Tibshirani 1996; Wu et al. 2009; Xu 2010), Elastic net (Zou and Hastie 2005; Cho et al. 2010), Bayesian lasso (Park and Casella 2008; Yi and Xu 2008) as well as adaptive Bayesian Lasso (Sun et al. 2010; Mutshinda and Sillanpää 2010). For a recent review in genetics, see Li and Sillanpää (2012). For multivariate lasso methods, see Yuan et al. (2007) and Liu et al. (2012).

Normality of residuals is commonly assumed in univariate and multivariate regression methods that are used to study phenotype–genotype associations. To make phenotypes (and residuals) more normal, data transformations are applied (see Xu and Hu 2010). A robust alternative, which is less sensitive to the outlying observations, is to assume that the residuals follow some non-normal distribution (e.g., t-distribution or Laplace distribution) which have heavier than normal tails. For the univariate case, see Lange et al. (1989) and Yang et al. (2009), and for the multivariate case, see Chen et al. (2014). For expression data application, see Purdom and Holmes (2005). It is well known that L_2-norm based likelihood arises from assuming normally distributed residuals. Therefore, a robust alternative is obtained if one fits the model and use the L_1-norm instead of the L_2-norm in place of the likelihood. This kind of approach is called LAD-lasso and has been presented for econometrics applications (Wang et al. 2007). A generalization of LAD-lasso to make it adaptive is called weighted LAD-lasso (Arslan 2012). Also lasso approaches combined with Huber's criterion (Lambert-Lacroix and Zwald 2011), minimum distance criterion (Chi and Scott 2014) or heteroscedastic sparse regression (Daye et al. 2012) as well as robust least angle regression (Khan et al. 2007) or the least trimmed squares (Alfons et al. 2013) have been proposed as alternatives for LAD-lasso approach. See also Zhou et al. (2013) and Gao and Huang (2010). However, here we adopt normal LAD-lasso and bring it to genomics context (cf. Daye et al. 2012) and generalize it for multivariate data. The robust performance of the method is illustrated by using multivariate simulated data examples.

14.2 Univariate Lasso and LAD-Lasso

Consider the univariate multiple regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ is an $n \times 1$ vector of n observed values, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ is an $n \times q$ matrix of n observed values of q explaining variables, $\boldsymbol{\beta}$ is a $q \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is an $n \times 1$ vector of residuals. We further assume that the model has an intercept term, i.e. the first column of $\mathbf{X}$ is a column of ones. The ordinary least squares estimate $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes the objective

function

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{14.1}$$

and when the rank of $\mathbf{X}$ is q, the minimizing value is

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

Tibshirani (1996) introduced a shrinkage estimation method for the univariate multiple regression. The ordinary least squares objective function is now minimized subject to $\sum_{j=2}^{q}|\beta_j| \leq t$ where $t \geq 0$ is a tuning parameter. It can be shown that minimization of (14.1) subject to $\sum_{j=2}^{q}|\beta_j| \leq t$ is equivalent to minimization of the objective function

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2 + \lambda\sum_{j=2}^{q}|\beta_j|, \tag{14.2}$$

where $\lambda \geq 0$ is a tuning parameter. The intercept β_1 is not included in the lasso penalty term since we are only interested in sparsity in the covariates. The minimizer of the objective function (14.2) is called a *lasso estimate* for $\boldsymbol{\beta}$. Note that the choice $\lambda = 0$ gives the usual ordinary least squares regression solution.

It is well known that the ordinary least squares estimation method (and consequently the lasso-estimation method) is very sensitive to outliers (Alfons et al. 2013). A robust alternative for the lasso-estimate is achieved by replacing the squared residuals with absolute values of residuals in the objective function (14.2). See, e.g., Wang et al. (2007), Wang and Leng (2007), Xu and Ying (2010), and Arslan (2012). The minimizer of the objective function

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \boldsymbol{\beta}^\top \mathbf{x}_i| + \lambda\sum_{j=2}^{q}|\beta_j| \tag{14.3}$$

is then the so-called *LAD-lasso estimate* for $\boldsymbol{\beta}$. Obviously the choice $\lambda = 0$ gives the usual LAD (least absolute deviations) regression solution. Let

$$\begin{pmatrix} y_i^* \\ \mathbf{x}_i^* \end{pmatrix} = \begin{cases} \begin{pmatrix} y_i \\ \mathbf{x}_i \end{pmatrix}, & \text{if } i = 1, \ldots, n, \\ \begin{pmatrix} 0 \\ n\lambda\mathbf{e}_{i-n+1} \end{pmatrix}, & \text{if } i = n+1, \ldots, n+q-1, \end{cases}$$

where $\mathbf{e}_i$ is a $q \times 1$ standard unit vector with a 1 as the ith component and zeroes for all the other components. Now the objective function (14.3) can be rewritten in the

form

$$\frac{1}{n+q-1}\sum_{i=1}^{n+q-1}|y_i^* - \boldsymbol{\beta}^\top \mathbf{x}_i^*|, \tag{14.4}$$

which implies that the LAD-lasso estimate can be found by using any standard LAD-estimation program, e.g. using function *rq* in the R-package *quantreg* or the *LAV* command in the SAS/IML library.

14.3 Multivariate Lasso and LAD-Lasso

Consider now multivariate multiple regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{14.5}$$

where $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^\top$ is an $n \times p$ matrix of n observed values of p response variables, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ is an $n \times q$ matrix of n observed values of q explaining variables, $\boldsymbol{\beta}$ is a $q \times p$ matrix of regression coefficients, and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)^\top$ is an $n \times p$ matrix of residuals. We further assume that $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ is a random sample of size n from a p-variate distribution centred at the origin.

The minimizer of the objective function

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{y}_i - \boldsymbol{\beta}^\top \mathbf{x}_i\|^2 &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^\top(\mathbf{y}_i - \boldsymbol{\beta}^\top \mathbf{x}_i)\\ &= \frac{1}{n}\operatorname{tr}\left[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right]\end{aligned} \tag{14.6}$$

gives the ordinary least squares estimate and when the rank of $\mathbf{X}$ is q, the minimizing value is

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}.$$

The univariate lasso estimation method can be generalized straightforwardly to the multivariate case by using the penalized objective function

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{y}_i - \boldsymbol{\beta}^\top \mathbf{x}_i\|^2 + \lambda\sum_{j=2}^{q}\|\boldsymbol{\beta}_j\|. \tag{14.7}$$

See, e.g., Turlach et al. (2005), Yuan and Lin (2006), and Yuan et al. (2007). The minimizer of the objective function (14.7) gives now *multivariate lasso estimate* for the regression coefficient matrix $\boldsymbol{\beta}$. Note that the model (14.5) can be rewritten as

$$\mathbf{y}_* = \mathbf{X}_*\boldsymbol{\beta}_* + \boldsymbol{\epsilon}_* = \left(\mathbf{X}_{*1}\cdots\mathbf{X}_{*q}\right)\boldsymbol{\beta}_* + \boldsymbol{\epsilon}_* = \sum_{j=1}^{q}\mathbf{X}_{*j}\boldsymbol{\beta}_j + \boldsymbol{\epsilon}_*, \tag{14.8}$$

where $\mathbf{y}_* = \left(\mathbf{y}_1{}^\top \cdots \mathbf{y}_n{}^\top\right)^\top$ is an $np \times 1$ vector, $\mathbf{X}_* = \mathbf{X} \otimes \mathbf{I}_p$ is an $np \times pq$ matrix, $\boldsymbol{\beta}_* = \left(\boldsymbol{\beta}_1{}^\top \cdots \boldsymbol{\beta}_q{}^\top\right)^\top$ is a $pq \times 1$ vector and $\boldsymbol{\epsilon}_* = \left(\boldsymbol{\varepsilon}_1{}^\top \cdots \boldsymbol{\varepsilon}_n{}^\top\right)^\top$ is $np \times 1$ vector. The Eq. (14.8) implies that the multivariate multiple regression model is of the same form as the regression problem with q factors where the $np \times p$ matrix $\mathbf{X}_{*j}$ corresponds to the jth factor and $\boldsymbol{\beta}_j$ is a coefficient vector (Yuan and Lin 2006). The penalized objective function for multivariate lasso (14.7) is therefore also an objective function for group lasso problem (Yuan and Lin 2006).

The multivariate lasso method gives sparse solutions but it is obviously not very robust. See Sect. 14.6 for the robustness study. In the same way as in the univariate case you get a more robust version by replacing the squared norms with norms. In other words, you minimize the penalized objective function

$$\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \boldsymbol{\beta}^\top \mathbf{x}_i\| + \lambda \sum_{j=2}^{q} \|\boldsymbol{\beta}_j\| \tag{14.9}$$

with respect to the coefficient matrix $\boldsymbol{\beta}$. Note that when using the penalty term $\lambda \sum_{j=2}^{q} \|\boldsymbol{\beta}_j\|$, we get a robustified version of group lasso objective function (14.7). Denote the minimizing value of (14.9) by $\hat{\boldsymbol{\beta}}_{LL}$. It is easily seen that if we define, in the same way as in Sect. 14.2,

$$\begin{pmatrix} \mathbf{y}_i^* \\ \mathbf{x}_i^* \end{pmatrix} = \begin{cases} \begin{pmatrix} \mathbf{y}_i \\ \mathbf{x}_i \end{pmatrix}, & \text{if } i = 1, \ldots, n, \\ \begin{pmatrix} \mathbf{0} \\ n\lambda \mathbf{e}_{i-n+1} \end{pmatrix}, & \text{if } i = n+1, \ldots, n+q-1, \end{cases}$$

the objective function (14.9) reduces to the LAD estimation objective function (Oja 2010)

$$\frac{1}{n+q-1} \sum_{i=1}^{n+q-1} \|\mathbf{y}_i^* - \boldsymbol{\beta}^\top \mathbf{x}_i^*\|, \tag{14.10}$$

which shows that we can use any multivariate LAD regression estimation routine to find the multivariate LAD-lasso estimate $\hat{\boldsymbol{\beta}}_{LL}$. You can, for example, use the function *mv.l1lm* of the R-package *MNM* (Nordhausen et al. 2009; Nordhausen and Oja 2011).

14.4 Determination of the Tuning Parameter

To choose the tuning parameter in the univariate case it has been proposed to use AIC, BIC, GCV criteria or to perform cross-validation (Hastie et al. 2009). In the multivariate case, some changes to the criteria may be needed. We have decided to

modify the BIC criterion to handle the multivariate LAD case. In our new BIC-type LAD (BTL) criterion, we have replaced the OLS objective function with the LAD objective function. Additionally, to calculate degrees of freedom in multivariate case, we have adopted the convention to calculate the number of non-zero regression vectors at each covariate position. We also assume that $\sigma^2 = \text{var}(\varepsilon_{ij})$. We use the following criterion to determine the tuning parameter λ:

$$BTL_\lambda = \frac{\sum_{i=1}^{n} \|\mathbf{y}_i - \boldsymbol{\beta}(\lambda)^\top \mathbf{x}_i\|}{n\,\sigma^2} + \frac{\log(n)}{n} h(\lambda), \tag{14.11}$$

where $\boldsymbol{\beta}(\lambda)_j$ is the jth row of the LAD-lasso estimate $\boldsymbol{\beta} = \boldsymbol{\beta}(\lambda)$ and $h(\lambda) = \sum_{j=2}^{q} I(\|\boldsymbol{\beta}(\lambda)_j\| > 0)$ is the number of non-zero regression vectors. We then choose the λ that minimizes the BTL-criterion, i.e.

$$\lambda_0 = \arg\min_{\lambda \in \Lambda} BTL_\lambda,$$

where $\Lambda = \{\lambda_{min}, \lambda_{min} + \delta, \lambda_{min} + 2\delta, \ldots, \lambda_{max}\}$, $\delta > 0$ is step length and $\lambda_{max} = \lambda_{min} + (M-1)\delta$.

14.5 Significance Testing

After LAD-lasso estimation of the effects, especially with a small λ value, it is useful to do hypothesis testing for each covariate separately to control false positive findings. For conventional lasso, Meinshausen et al. (2009) suggested a sample-split method where data over individuals are randomly divided into two equal-sized samples. The first sample is used for estimation of the set of active predictors (i.e. the set $S = \{j; \beta_j \neq 0\}$) and the second sample is used for fitting the active predictors with OLS. The p-values $\tilde{P}_j, j \in S$, are then calculated for the predictors of the model and for the noise variables ($S^c = \{j; \beta_j = 0\}$) the p-values are set to 1. The adjusted p-values are defined as $\check{P}_j = \min(|S|\tilde{P}_j,\ 1), j = 1, \ldots, q$. These steps are repeated B times and after that aggregated p-values are calculated:

$$P_j = \min\left\{1, (1 - \log(0.05)) \inf_{\gamma \in (0.05,1)} Q_j(\gamma)\right\},$$

where $Q_j(\gamma) = \min\{1, q_\gamma(\{\check{P}_j^{(b)}/\gamma; b = 1, \ldots, B\})\}$, $\check{P}_j^{(b)}$ is the p-value of the jth variable of the bth split and $q_\gamma(\cdot)$ is the empirical γ-quantile function. Finally, variable j is selected if $P_j \leq \alpha$. This same procedure can be applied also for the univariate and multivariate LAD-lasso by fitting the active predictors with LAD and calculating the p-values with the function *mv.l1lm* of the R-package *MNM* (Nordhausen et al. 2009; Nordhausen and Oja 2011).

14.6 A Simulation Study

For this study, we simulated new phenotypes with trivariate traits using the public genotype dataset from the 12th QTL-MAS Workshop (QTL-MAS 2008). For more information on the dataset, see Crooks et al. (2009) and Lund et al. (2009). The original genotype dataset contains 5865 individuals and 6000 markers. To reduce rank-deficiency and collinearity among explanatory variables, we took a random sample of 300 from the individuals and chose every 30th marker with a resulting total number of 200 markers. We then generated trivariate traits by using the multivariate multiple regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y}$ is a 300×3 matrix of trivariate traits, $\mathbf{X}$ is a 300×200 matrix with ijth element

$$x_{ij} = \begin{cases} -1, \text{ if indiv. } i \text{ is homozygote (11) at marker } j, \\ 0, \text{ if indiv. } i \text{ is heterozygote (12) at marker } j, \\ 1, \text{ if indiv. } i \text{ is homozygote (22) at marker } j, \end{cases}$$

$\boldsymbol{\beta}$ is a 200×3 matrix with four QTLs indicated as non-zero rows

$$\boldsymbol{\beta}_{50}{}^{\top} = (100, 100, 100), \quad \boldsymbol{\beta}_{75}{}^{\top} = (0, 50, 100)$$
$$\boldsymbol{\beta}_{100}{}^{\top} = (5, 10, 15) \text{ and } \boldsymbol{\beta}_{150}{}^{\top} = (3, 3, 3),$$

and $\boldsymbol{\varepsilon}$ is a 300×3 matrix with i.i.d. rows distributed as

$$N_3\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}1.0 & 0.5 & 0.3\\0.5 & 1.0 & 0.2\\0.3 & 0.2 & 1.0\end{pmatrix}\right).$$

We estimated the tuning parameter λ by using *BTL*-criterion. Figure 14.1 shows that *BTL* has minimum value at $\lambda = 0.105$. We then estimated $\boldsymbol{\beta}$ using both multivariate lasso and multivariate LAD-lasso method. Figure 14.2 shows that when there are no outliers, multivariate lasso and multivariate LAD-lasso methods perform equally well for finding the four QTLs. We then generated a contaminated data set by multiplying $\mathbf{y}_{10}$ and $\mathbf{y}_{292}$ by 100:

$$\mathbf{y}_{10} = (-8.3, 36.9, 82.1) \rightarrow \mathbf{y}_{10} = (-834.5, 3689.5, 8211.8)$$

and

$$\mathbf{y}_{292} = (-5.9, 39.6, 85.2) \rightarrow \mathbf{y}_{292} = (-588.8, 3962.2, 8519.1).$$

Figure 14.1 shows that *BTL* has minimum value when the tuning parameter is $\lambda = 0.105$. We can clearly see that the outliers do not have much effect on

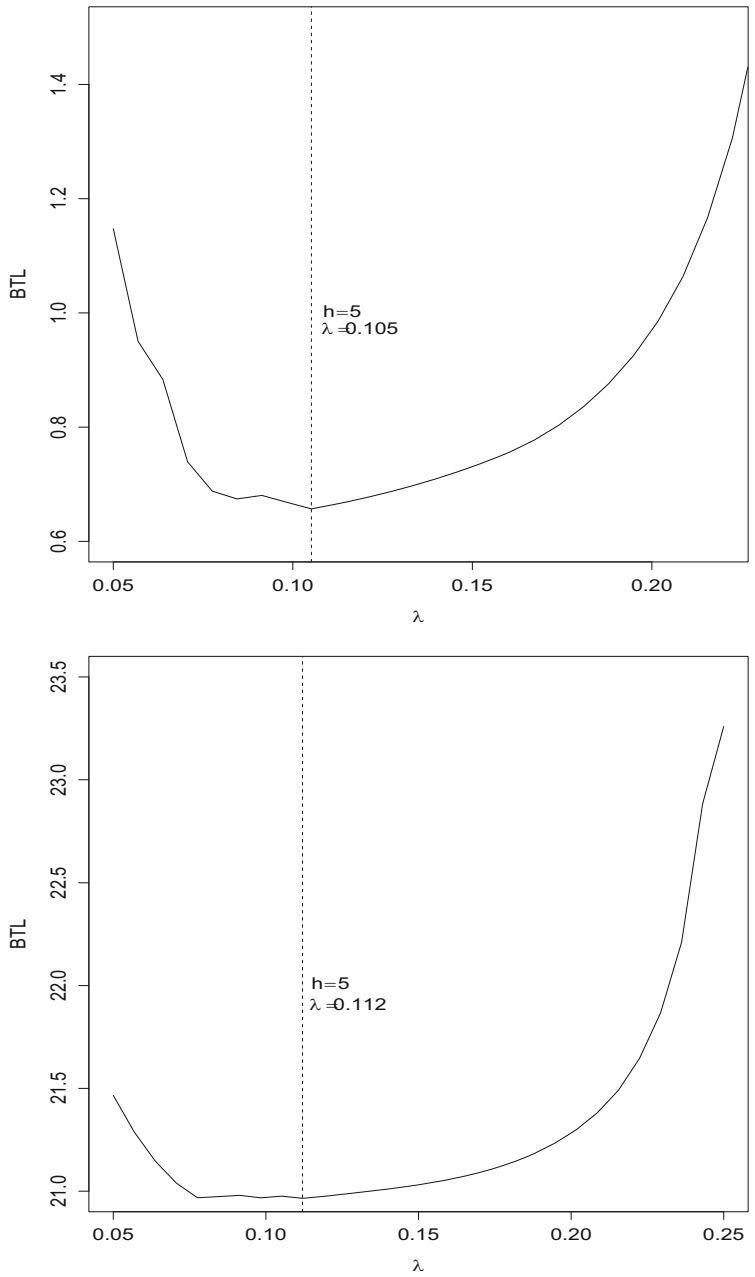

Fig. 14.1 *BTL*-criterion as a function of tuning parameter λ for the original data (*top*) and the contaminated data (*bottom*). The parameter h gives the number of non-zero regression vectors $\boldsymbol{\beta}_j$ including the intercept

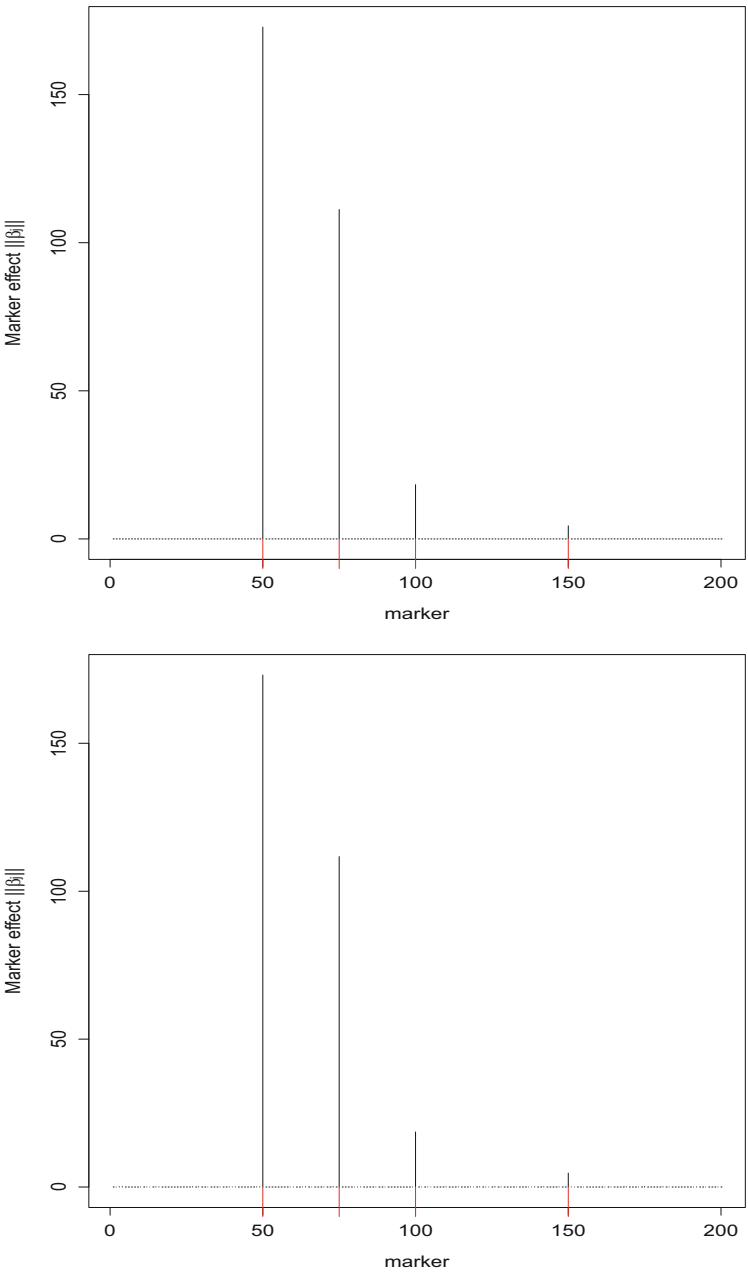

Fig. 14.2 Marker effects $\|\boldsymbol{\beta}_j\|$, $j = 1, \ldots 200$, for multivariate LAD-lasso estimate (*top*) and multivariate lasso estimate (*bottom*) for the original data using tuning parameter $\lambda = 0.105$

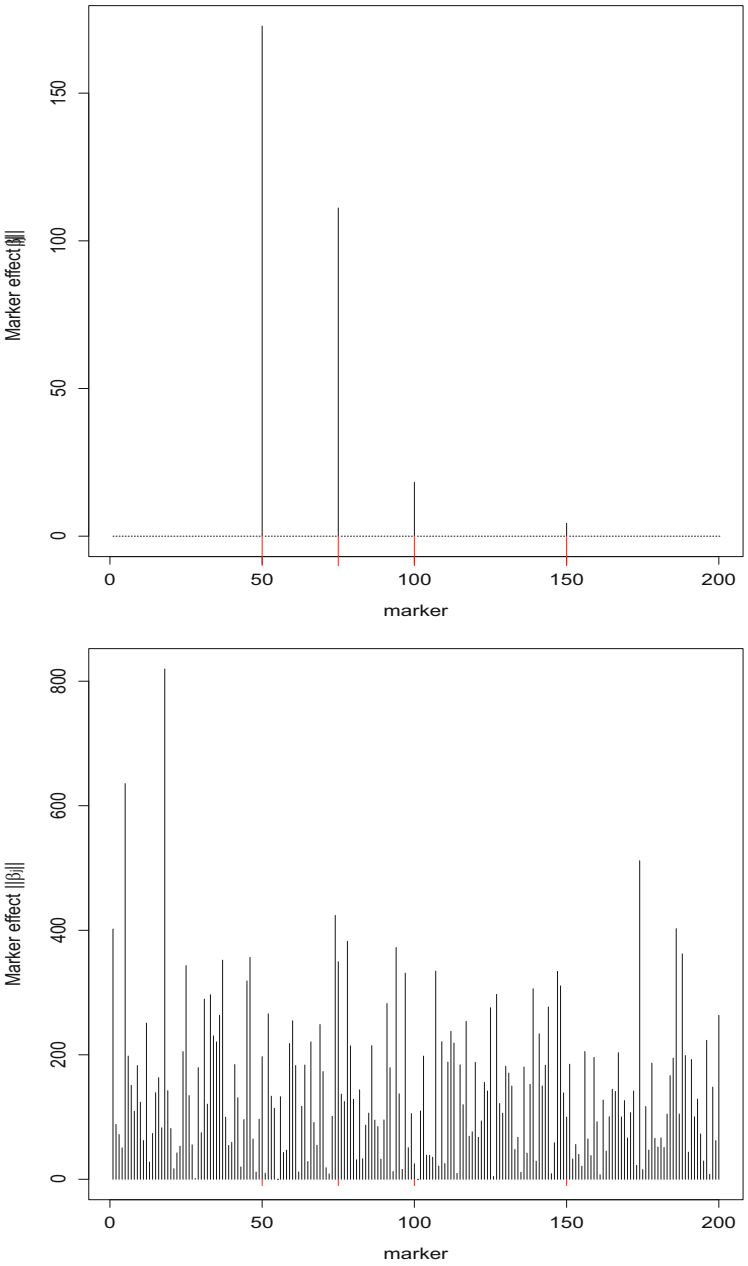

Fig. 14.3 Marker effects $\|\boldsymbol{\beta}_j\|$, $j = 1, \ldots 200$, for multivariate LAD-lasso estimate (*top*) and multivariate lasso estimate (*bottom*) for the contaminated data using tuning parameter $\lambda = 0.112$

the multivariate LAD-lasso (see Fig. 14.3, top) but they have drastic influence on multivariate lasso (see Fig. 14.3, bottom).

14.7 Concluding Remarks

Univariate and multivariate lasso estimation methods are highly sensitive to outlying observations because of the sum of squared norms term in the objective function. Using sum of norms instead of sum of squared norms gives us a considerably more robust estimate for the regression coefficients. In the example analyses, we analysed well-behaving subsamples where rank-deficiency and collinearity were fully controlled but phenotype distribution contained outlying observations. This was done in order to better see the robustifying influence of LAD to the outlying observations in the response variable. LAD-regression method and obviously also the multivariate LAD-lasso method are not robust against leverage points (outliers with respect to the explanatory variables) but it should not be a problem when the explanatory variables are discrete (as in our simulation study here).

It is important to emphasize here that lasso regression as well as our method has been designed especially to the cases where there are more explanatory variables than observations (rank-deficiency). Also, in genetic problems there are often substantial collinearity among explanatory variables. More extensive testing of our method in presence of more realistic genetic data will be done elsewhere (Li et al. 2015). Concerning scalability, LAD-lasso approach presented here can be easily used to analyse few hundreds of markers and of individuals within a reasonable time. However, the execution time of LAD-lasso R-programs is clearly slower than corresponding estimation programs which use lasso coordinate descent algorithm. Better estimation algorithms for LAD-lasso would be needed for rapid LAD-lasso analysis in the future.

References

Alfons, A., Croux, C., Gelper, S.: Sparse least trimmed squares regression for analyzing high-dimensional large data sets. Ann. Appl. Stat. **7**, 226–248 (2013)

Arslan, O.: Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression. Comput. Stat. Data Anal. **56**, 1952–1965 (2012)

Chen, L., Pourahmadi, M., Maadooliat, M.: Regularized multivariate regression models with skew-t error distributions. J. Stat. Plan. Inference **149**, 125–139 (2014)

Chi, E.C., Scott, D.W.: Robust parametric classification and variable selection by a minimum distance criterion. J. Comput. Graph. Stat. **23**, 111–128 (2014)

Cho, S., Kim, K., Kim, Y.J., Lee, J.-K., Cho, Y.S., Lee, J.-Y., Han, B.-C., Kim, H., Ott, J., Park, T.: Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. Ann. Hum. Genet. **74**, 416–428 (2010)

Crooks, L., Sahana, G., de Koning, D.J., Lund, M.S., Carlborg, Ö.: Comparison of analyses of the QTLMAS XII common dataset. II: genome-wide association and fine mapping. BMC Proc. **3**(Suppl 1), S2 (2009)

Daye, Z.J., Chen, J., Li, H.: High-dimensional heteroscedastic regression with an application to eQTL data analysis. Biometrics **68**, 316–326 (2012)

Gao, X., Huang, J.: Asymptotic analysis of high-dimensional LAD-regression with LASSO. Stat. Sin. **20**, 1485–1506 (2010)

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Prediction, Inference and Data Mining, 2nd edn. Springer, New York (2009)

Khan, J.A., Van Aelst, S., Zamar, R.H.: Robust linear model selection based on least angle regression. J. Am. Stat. Assoc. **102**, 1289–1299 (2007)

Lambert-Lacroix, S., Zwald, L.: Robust regression through the Huber's criterion and adaptive lasso penalty. Electron. J. Stat. **5**, 1015–1053 (2011)

Lange, K.L., Little, R.J.A., Taylor, J.M.G.: Robust statistical modeling using the t distribution. J. Am. Stat. Assoc. **84**, 881–896 (1989)

Li, Z., Sillanpää, M.J.: Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theor. Appl. Genet. **125**, 419–435 (2012)

Li, Z., Möttönen, J., Sillanpää, M.J.: A robust multiple-locus method for quantitative trait locus analysis of non-normally distributed multiple traits. Heredity (2015, in press)

Liu, J., Huang, J., Ma, S.: Analysis of genome-wide association studies with multiple outcomes using penalization. PLoS One **7**, e51198 (2012)

Lund, M.S., Sahana, G., de Koning, D.J., Su, G., Carlborg, Ö.: Comparison of analyses of the QTLMAS XII common dataset, I. Genomic selection. BMC Proc. **3**(Suppl 1), S1 (2009)

Meinshausen, N., Meier, L., Bühlmann, P.: p-values for high-dimensional regression. J. Am. Stat. Assoc. **104**, 1671–1681 (2009)

Mutshinda, C.M., Sillanpää, M.J.: Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics **186**, 1067–1075 (2010)

Nordhausen, K., Oja, H.: Multivariate L_1 methods: the package MNM. J. Stat. Softw. **43**, 1–28 (2011)

Nordhausen, K., Möttönen, J., Oja, H.: MNM: Multivariate Nonparametric Methods. An Approach Based on Spatial Signs and Ranks. R package version 0.95-1. http://CRAN.R-project.org/package=MNM (2009)

O'Hara, R.B., Sillanpää, M.J.: Review of Bayesian variable selection methods: what, how and which. Bayesian Anal. **4**, 85–118 (2009)

Oja, H.: Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks. Lecture Notes in Statistics, vol. 199. Springer, Heidelberg (2010)

Park, T., Casella, G.: The Bayesian LASSO. J. Am. Stat. Assoc. **103**, 681–686 (2008)

Purdom, E., Holmes, S.P.: Error distribution for gene expression data. Stat. Appl. Genet. Mol. Biol. **4**(1), article 16 (2005)

QTL-MAS: Data [online]. Available at: http://www.computationalgenetics.se/QTLMAS08/QTLMAS/DATA.html (2012). Cited 8 Feb 2012

Sun, W., Ibrahim, J.G., Zou, F.: Genome-wide multiple loci mapping in experimental crosses by the iterative adaptive penalized regression. Genetics **185**, 349–359 (2010)

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**, 267–288 (1996)

Turlach, B.A., Venables, W.N., Wright, S.J.: Simultaneous variable selection. Technometrics **47**, 349–363 (2005)

Wang, H., Leng, C.: Unified lasso estimation by least squares approximation. J. Am. Stat. Assoc. **102**, 1039–1048 (2007)

Wang, H., Li, G., Jiang, G.: Robust regression shrinkage and consistent variable selection through the LAD-lasso. J. Bus. Econ. Stat. **25**, 347–355 (2007)

Wu, T.T., Chen, F., Hastie, T., Sobel, E., Lange, K.: Genome-wide association analysis by lasso penalized regression. Bioinformatics **25**, 714–721 (2009)

Xu, S.: An expectation-maximization algorithm for the Lasso estimation of quantitative trait locus effects. Heredity **105**, 483–494 (2010)

Xu, S., Hu, Z.: Generalized linear model for interval mapping of quantitative trait loci. Theor. Appl. Genet. **121**, 47–63 (2010)

Xu, J., Ying, Z.: Simultaneous estimation and variable selection in median regression using Lasso-type penalty. Ann. Inst. Stat. Math. **62**, 487–514 (2010)

Yang, R., Wang, X., Li, J., Deng, H.: Bayesian robust analysis for genetic architecture of quantitative traits. Bioinformatics **25**(8), 1033–1039 (2009)

Yi, N., Xu, S.: Bayesian LASSO for quantitative trait loci mapping. Genetics **179**, 1045–1055 (2008)

Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B **68**, 49–67 (2006)

Yuan, M., Ekici, A., Lu, Z., Monteiro, R.: Dimension reduction and coefficient estimation in multivariate linear regression. J. R. Stat. Soc. Ser. B **69**, 329–346 (2007)

Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B **67**, 301–320 (2005)

Zhou, Z., Jiang, R., Qian, W.: LAD variable selection for linear models with randomly censored data. Metrika **76**, 287–300 (2013)

Chapter 15
On Some Nonparametric Classifiers Based on Distribution Functions of Multivariate Ranks

Olusola Samuel Makinde and Biman Chakraborty

Abstract Over the last two decades, multivariate sign and rank based methods have become popular in analysing multivariate data. In this paper, we propose a classification methodology based on the distribution of multivariate rank functions. The proposed method is fully nonparametric in nature. Initially, we consider a theoretical version of the classifier for K populations and show that it is equivalent to the Bayes rule for spherically symmetric distributions with a location shift. Then we present the empirical version of that and show that the apparent misclassification rate of the empirical version of the classifier converges asymptotically to the Bayes risk. We also present an affine invariant version of the classifier and its optimality for elliptically symmetric distributions. We illustrate the performance in comparison with some other depth based classifiers using simulated and real data sets.

Keywords Bayes risk • Elliptically symmetric distributions • Location shift • Maximal depth • Spatial outlyingness

15.1 Introduction

With the growth of internet, a huge amount of online information is available these days and it has become paramount to almost all businesses to use these information to target their potential customers. In such applications classification has become one of the most widely used tools to statisticians as well as computer scientists. The objective in a classification problem is to classify an observation into one of several competing classes or populations. If the prior probabilities of these populations are given by p_i and the probability density functions are given by f_is, then the Bayes rule classifies an observation $\mathbf{x}$ to the population j, which has the maximum posterior probability ($\propto p_j f_j(\mathbf{x})$) at $\mathbf{x}$. The Bayes rule has the lowest possible expected misclassification error rate. However, in practice the probability density functions f_is are unknown, and they need to be estimated using

O.S. Makinde • B. Chakraborty (✉)
School of Mathematics, University of Birmingham, Birmingham B15 2TT, UK
e-mail: OSM151@student.bham.ac.uk; B.Chakraborty@bham.ac.uk

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_15

the training sample. Commonly used parametric methods like linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) (see Fisher 1936) are motivated by specific distributional assumptions like multivariate normality of the populations. Alternatively, if we restrict to a linear classification rule, we may like to determine a hyperplane $\mathbf{a}^\top \mathbf{x} + b = 0$, which separates the populations best. There are several methods for choosing such a hyperplane based on the training sample available in the literature (Fukunaga 1990; McLachlan 1992; Hastie et al. 2001). Similarly, one can also look for quadratic separating surfaces. Fisher's LDA and QDA can also be viewed as linear and quadratic separating surfaces, where those surfaces were estimated using the mean vector and the covariance matrices of the training samples. Naturally, LDA and QDA are very sensitive to outliers if there are any in the training samples as they depend on the sample moment estimators. The linear and quadratic surface based classifiers may lead to poor classification if the separating surfaces are not linear or quadratic. For these reasons, nonparametric classifiers, which are more flexible to distributional assumptions and extreme values, are more desirable.

In the last couple of decades, notions of multivariate signs and ranks have become a useful tool in analysing multivariate data, as it does not depend on distributional assumptions heavily and characterizes the central and extreme observations quite effectively. Möttönen and Oja (1995), Möttönen et al. (1997) used the notion of spatial ranks to construct multivariate tests of location. Serfling (2010) presented a very detailed study on spatial rank functions and their equivariance and invariance properties. A related notion to multivariate ranks is the data depth. Vardi and Zhang (2000) proposed l_1-depth (or, spatial depth) related to the spatial ranks. Liu et al. (1999) proposed various ideas on analysing multivariate data using data depths. Christmann and Rousseeuw (2001) and Christmann et al. (2002) applied the idea of regression depth (Rousseeuw and Hubert 1999) to classification. Ghosh and Chaudhuri (2005a) used half-space depth and regression depth to construct linear and nonlinear separating curves or surfaces. Ghosh and Chaudhuri (2005b) proposed a classifier based on the notion of maximum depth. This rule was intuitively appealing and performs well when the populations differ only in location and the prior probabilities of the populations are equal. Jörnsten (2004) considered spatial depth to construct classifiers, which are different from maximal depth classifiers. Cui et al. (2008) considered a maximum depth classifier based on projection depth. Recently, Dutta and Ghosh (2012a,b) proposed classifiers based on projection depths and L_p-depths. Using the idea of DD-plots (Liu et al. 1999), Li et al. (2012) proposed DD-classifier, which is a modified version of maximum depth classifier. Lange et al. (2014) proposed another classifier based on zonoid depths. Recently, Paindaveine and Van Bever (2015) proposed another depth based classifier which uses the notions of local depth. However, many of these methods based on data depths suffer from computational complexities of the depth functions.

In this paper, we define an outlyingness function based on spatial ranks. This notion of multivariate outlyingness is closely related to the depth functions. Makinde (2014) considered a classifier on minimizing that spatial outlyingness function, which is equivalent to maximum depth classifiers. Here we propose a

fully nonparametric classifier based on the distribution functions of the spatial outlyingness functions. The method is fully nonparametric in nature and computationally easy. We show that the proposed classifier is equivalent to the Bayes rule for spherically symmetric distributions with a location shift only. We also propose an affine invariant version of the classifier, which is equivalent to the Bayes rule for elliptically symmetric distribution. We compare the performance of the proposed classifier with LDA, QDA, and some depth based classifiers. We also illustrate the performance using some benchmark real data sets.

15.2 Classifier Based on Distribution of Multivariate Outlyingness

Suppose $\mathbf{X}$ is a d-dimensional random vector (or, feature vector) having a distribution F, which is assumed to be absolutely continuous with respect to the Lebesgue measure $\mathbb{R}^d$. The spatial rank function (Möttönen and Oja 1995) of any point $\mathbf{x} \in \mathbb{R}^d$ with respect to F is defined as

$$rank_F(\mathbf{x}) = E_F\left[\frac{\mathbf{x}-\mathbf{X}}{\|\mathbf{x}-\mathbf{X}\|}\right]. \tag{15.1}$$

Here $\|\cdot\|$ is the usual Euclidian norm. It follows immediately from the definition that $rank_F(\mathbf{x}) = 0$ implies that $\mathbf{x}$ is the spatial median of the multivariate distribution F. Koltchinskii (1997) established that this spatial rank function is a one-to-one function of the distribution function F and hence it characterizes the distribution.

Define $r_F(\mathbf{x}) = \|rank_F(\mathbf{x})\|$ and then define the distribution function of $r_F(\mathbf{X})$ as

$$F_R(r) = P\left(\|rank_F(\mathbf{X})\| \le r\right), \tag{15.2}$$

where $\mathbf{X}$ has distribution F. Liu and Singh (1993), Liu (1995) used the distribution function of the data depths to construct quality index and multivariate control charts. This distribution of the multivariate rank function has been used earlier to construct the central rank regions and some descriptive measures like scale curves based on volume functionals in Guha and Chakraborty (2013) and Serfling (2002). It is easy to observe that $F_R(r_F(\mathbf{x}))$ gives an idea on how outlying $\mathbf{x}$ is with respect to the distribution F. If $\mathbf{x}$ is a central observation, $r_F(\mathbf{x})$ will be small and consequently, $F_R(r_F(\mathbf{x}))$ will be small. On the other hand, $\mathbf{x}$ is an extreme observation, then $F_R(r_F(\mathbf{x}))$ will be large. Thus a large value of $F_R(r_F(\mathbf{x}))$ may suggest a deviation of $\mathbf{x}$ from the population π_1. We can also observe the following simple fact:

Lemma 15.1 *Let* $\mathbf{X}$ *be a* d*-dimensional random vector having distribution* F*, which is absolutely continuous with respect to the Lebesgue measure in* $\mathbb{R}^d$*. Then* $F_R(r_F(\mathbf{X}))$ *will have a uniform*$(0, 1)$ *distribution.*

Now we propose a multivariate classifier based on this distribution function of the multivariate ranks. Consider K classes $\pi_1, \ldots, \pi_K$ with distribution functions

$F_1, \ldots, F_K$, respectively. Let c_{ij} be the cost of misclassifying an observation in π_j when it is actually from π_i, for $i, j = 1, \ldots, K$. Let us assume that all c_{ij}'s are equal and prior probabilities of the classes $p_1 = p_2 = \cdots = p_K$. Then using the motivation stated above, we define a classification rule for any $\mathbf{x} \in \mathbb{R}^d$ as

$$\text{assign } \mathbf{x} \text{ to } \pi_j \text{ if } F_{R_j}(r_{F_j}(\mathbf{x})) = \min_{1 \le i \le K} F_{R_i}(r_{F_i}(\mathbf{x})), \tag{15.3}$$

where F_{R_i} is the distribution function corresponding to the outlyingness function $\|rank_{F_i}(\mathbf{X}_i)\|$ with $\mathbf{X}_i$ having distribution function F_i, for $i = 1, \ldots, K$.

If the classes $\pi_j, j = 1, \ldots, K$ have spherically symmetric distributions separated by location shifts only, we can show that the above classifier is equivalent to the Bayes rule and hence optimal. The result is more formally stated in the following:

Theorem 15.1 *Let F_j be spherically symmetric distributions with density functions of the form*

$$f_j(\mathbf{x}) = h((\mathbf{x} - \boldsymbol{\theta}_j)^\top (\mathbf{x} - \boldsymbol{\theta}_j)), \quad \mathbf{x} \in \mathbb{R}^d,$$

$j = 1, \ldots, K$, for some strictly decreasing, continuous, non-negative scalar function h. Then the classifier defined in (15.3) is equivalent to the Bayes rule when the prior probabilities $p_1 = p_2 = \cdots = p_K = 1/K$.

Proof For the prior probabilities $p_1 = p_2 = \cdots = p_K = 1/K$, the Bayes classification rule is given by

$$\text{assign } \mathbf{x} \text{ to } \pi_j \text{ if } f_j(\mathbf{x}) = \max_{1 \le j \le K} f_i(\mathbf{x}).$$

Under the assumptions of spherical symmetry for F_j, we observe that (Möttönen et al. 1997; Marden 1998) for $j = 1, \ldots, K$,

$$rank_{F_j}(\mathbf{x}) = h_1(\|\mathbf{x} - \boldsymbol{\theta}_j\|) \frac{\mathbf{x} - \boldsymbol{\theta}_j}{\|\mathbf{x} - \boldsymbol{\theta}_j\|}$$

for some monotonically increasing function h_1. This implies $r_{F_j}(\mathbf{x}) = h_1(\|\mathbf{x} - \boldsymbol{\theta}_j\|)$. Then by the assumption on the density functions of F_js, the functions $F_{R_j}(r))$ are identical for all $j = 1, \ldots, K$. Let us denote it by $F_R(r)$ and since $F_R(r_{F_j}(\mathbf{x}))$ are continuous increasing functions of $r_{F_j}(\mathbf{x})$, we can write $\|\mathbf{x} - \boldsymbol{\theta}_j\| = h_2(F_R(r_{F_j}(\mathbf{x})))$ for all $j = 1, \ldots, K$ for some increasing function h_2. Hence, the density function $f_j(\mathbf{x})$ is a decreasing function of $F_R(r_{F_j}(\mathbf{x}))$. So, the condition $f_j(\mathbf{x}) = \max_{1 \le i \le K} f_i(\mathbf{x})$ is equivalent to $F_R(r_{F_j}(\mathbf{x})) = \min_{1 \le i \le K} F_R(r_{F_i}(\mathbf{x}))$ and that proves that the proposed classifier in (15.3) is equivalent to the Bayes rule. □

In practice, the rank functions $rank_{F_j}$s will hardly be known completely and we need to estimate them from the training sample. Let $\mathbf{X}_{j,1}, \ldots, \mathbf{X}_{j,n_j} \in \mathbb{R}^d$ be a random sample from the population π_j having distribution function F_j for $j = 1, \ldots, K$. We

define the empirical rank functions as

$$rank_{F_{j,n_j}}(\mathbf{x}) = \frac{1}{n_j}\sum_{i=1}^{n_j}\frac{\mathbf{x}-\mathbf{X}_{j,i}}{\|\mathbf{x}-\mathbf{X}_{j,i}\|} \tag{15.4}$$

for any $\mathbf{x} \in \mathbb{R}^d$. Similarly, we can define the outlyingness functions as $r_{F_{j,n_j}}(\mathbf{x}) = \|rank_{F_{j,n_j}}(\mathbf{x})\|$ for $j = 1, \ldots, K$, respectively. The empirical version of F_{R_j}, denoted by $\hat{F}_{R_j}$, is the empirical distribution function of $r_{F_{j,n_j}}(\mathbf{X}_{j,1}), \ldots, r_{F_{j,n_j}}(\mathbf{X}_{j,n_j})$. Then the empirical classification rule for any $\mathbf{x} \in \mathbb{R}^d$ can be defined as

$$\text{assign } \mathbf{x} \text{ to } \pi_j \text{ if } \hat{F}_{R_j}(r_{F_{j,n_j}}(\mathbf{x})) = \min_{1\le i\le K}\hat{F}_{R_i}(r_{F_{i,n_i}}(\mathbf{x})), \tag{15.5}$$

For later reference, let us call the classifier defined in (15.5), rank distribution based classifier or RDC. For the large sample properties of the empirical version of the classifier, we observe the following result:

Lemma 15.2 *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *be independent and identically distributed d-dimensional random vectors having distribution function F, which is absolutely continuous, then*

$$\sup_{\mathbf{x}\in\mathbb{R}^d}\left|\hat{F}_R(r_{F_n}(\mathbf{x})) - F_R(r_F(\mathbf{x}))\right| \to 0 \quad a.s. \quad as\ n \to \infty.$$

Proof By Theorem 5.5 of Koltchinskii (1997),

$$\sup_{\mathbf{x}\in\mathbb{R}^d}|\|rank_{F_n}(\mathbf{x})\| - \|rank_F(\mathbf{x})\|| \to 0 \quad a.s.$$

as $n \to \infty$. Hence following the Lemma 3.1.2 of Guha (2012), we have

$$\sup_{\mathbf{x}\in\mathbb{R}^d}|\hat{F}_R(r_{F_n}(\mathbf{x})) - \hat{F}_R(r_F(\mathbf{x}))| \to 0 \quad a.s.$$

as $n \to \infty$. Also, by Glivenko–Cantelli lemma (Durrett 2010), $\sup_{r\in\mathbb{R}}|\hat{F}_R(r) - F_R(r)|$ converges to 0 almost surely as $n \to \infty$. The proof is now complete, by observing the fact that

$$\left|\hat{F}_R(r_{F_n}(\mathbf{x})) - F_R(r_F(\mathbf{x}))\right| \le \left|\hat{F}_R(r_{F_n}(\mathbf{x})) - \hat{F}_R(r_F(\mathbf{x}))\right| + \left|\hat{F}_R(r_F(\mathbf{x})) - F_R(r_F(\mathbf{x}))\right|.$$

□

15.3 Estimation of Classification Error

One way of evaluating the performance of a classification rule is to calculate its "error rates", or misclassification probabilities. One can define the *total probability of misclassification* (TPM) as

$$TPM = \sum_{j=1}^{K} p_j \int_{R_j^c} f_j(\mathbf{x}) d\mathbf{x}$$

where f_j is the probability density function of F_j and R_js denote the classification regions as follows:

$$R_j = \{\mathbf{x} \in \mathbb{R}^d : F_{R_j}(r_{F_j}(\mathbf{x})) = \min_{1 \le i \le K} F_{R_i}(r_{F_i}(\mathbf{x}))\}$$

for $j = 1, \ldots, K$. The empirical versions of the classification regions, $\hat{R}_j$s are given as:

$$\hat{R}_j = \{\mathbf{x} \in \mathbb{R}^d : \hat{F}_{R_j}(r_{F_{j,n_j}}(\mathbf{x})) = \min_{1 \le i \le K} \hat{F}_{R_i}(r_{F_{i,n_i}}(\mathbf{x}))\}$$

for $j = 1, \ldots, K$. So the *actual error rate* (AER) is given by

$$AER = \sum_{j=1}^{K} p_j \int_{\hat{R}_j^c} f_j(\mathbf{x}) d\mathbf{x}.$$

In real data problems, we cannot compute AER either as we do not know f_js completely and we need to estimate AER. It can be estimated using cross-validation (Lachenbruch and Mickey 1968). However, an easier way to evaluate a classifier, which does not depend on the parent populations, is to use the *apparent error rate* (APER). For a training sample $\mathbf{X}_{j,1}, \ldots, \mathbf{X}_{j,n_j} \in \pi_j$, $1 \le j \le K$ with $N = n_1 + \cdots + n_K$, the APER, $\hat{\Delta}_N$ can be defined as follows for our proposed classifier:

$$\hat{\Delta}_N = \sum_{j=1}^{K} \frac{p_j}{n_j} \sum_{i=1}^{n_i} I\{\mathbf{X}_{j,i} \notin \hat{R}_j\} \tag{15.6}$$

$\hat{\Delta}_N$ is easy to calculate and intuitively appealing, however, it is well known that it underestimates the actual error rate, AER, unless the sample sizes n_js are very large. The following result provides the asymptotic behaviour of the $\hat{\Delta}_N$.

Theorem 15.2 *Assume* $\min_{1\le j\le K} n_j \to \infty$ *such that* $n_j/N \to 1/K$, $1 \le j \le N$. *Then under the assumptions of Theorem 15.1 on the distributions* F_j *and* $p_1 = \cdots = p_K = 1/K$,

$$\hat{\Delta}_N \to \Delta_B \quad a.s.$$

where Δ_B *is the Bayes risk.*

Proof Let $\mathbf{X}_{j,1}, \ldots, \mathbf{X}_{j,n_j}$ be a random sample from the population π_j for $j = 1, \ldots, K$. Define

$$\tilde{\Delta}_N = \sum_{j=1}^{K} \frac{p_j}{n_j} \sum_{i=1}^{n_j} I\left\{\mathbf{X}_{j,i} \notin R_j\right\}.$$

By strong law of large numbers,

$$\frac{1}{n_j} \sum_{i=1}^{n} I\left\{\mathbf{X}_{j,i} \notin R_j\right\} \to \int_{R_j^c} f_j(\mathbf{x}) d\mathbf{x}$$

almost surely as $n_j \to \infty$. Note that the Bayes risk is given by

$$\Delta_B = \sum_{j=1}^{K} p_j \int_{R_j^c} f_j(\mathbf{x}) d\mathbf{x}$$

by Theorem 15.1. Hence

$$\tilde{\Delta}_N - \Delta_B \to 0 \quad a.s.$$

as $\min(n_1, \ldots, n_K) \to \infty$.

By the uniform almost sure convergence in Lemma 15.2,

$$\frac{1}{n_j} \sum_{i=1}^{n_j} I\left\{\mathbf{X}_{i,j} \notin \hat{R}_j\right\} - \frac{1}{n_j} \sum_{i=1}^{n} I\left\{\mathbf{X}_{i,j} \notin R_j\right\} \to 0 \quad a.s.$$

as $n_j \to \infty$. Hence as $\min(n_1, \ldots, n_K) \to \infty$, we have

$$\hat{\Delta}_N - \tilde{\Delta}_N \to 0 \quad a.s.$$

The proof is now complete by the triangle inequality

$$|\hat{\Delta}_N - \Delta_B| \le |\hat{\Delta}_N - \tilde{\Delta}_N| + |\tilde{\Delta}_N - \Delta_B|.$$

□

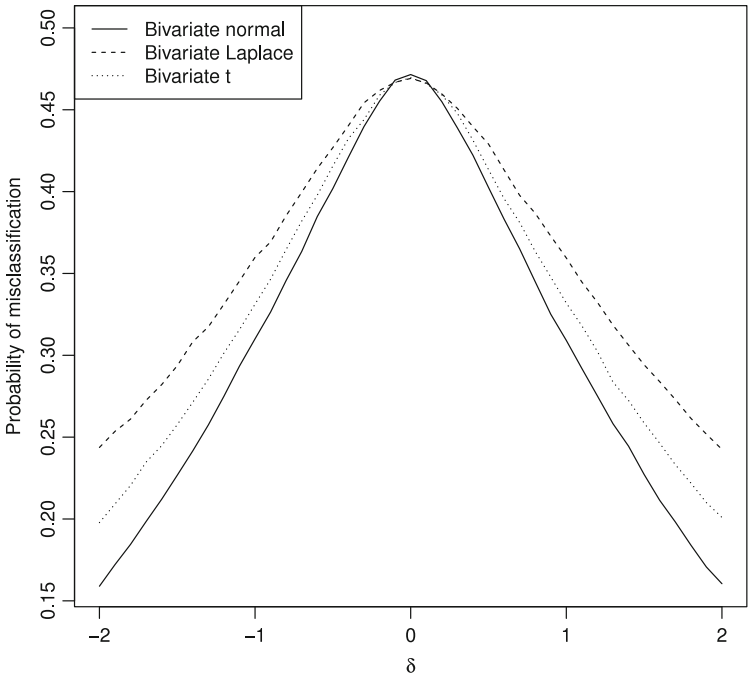

Fig. 15.1 AER for three different spherically symmetric distribution families with a location shift only

As illustration of actual error rates of our proposed empirical classifier in (15.5) (RDC), we present a small simulation study. Let us consider the two populations π_1 and π_2 to be bivariate spherically symmetric with centre of symmetries $\boldsymbol{\mu}_1 = (0,0)^\top$ and $\boldsymbol{\mu}_2 = (\delta,0)^\top$, respectively. The sample sizes for $\mathbf{X}_{1,1},\ldots,\mathbf{X}_{1,n_1}$ and $\mathbf{X}_{1,2},\ldots,\mathbf{X}_{2,n_2}$ are taken to be $n_1 = n_2 = 100$. We simulate a new random sample $\mathbf{Z}_1,\ldots,\mathbf{Z}_m$ from π_1 and $\mathbf{Z}_{m+1},\ldots,\mathbf{Z}_{2m}$ from π_2 with $m = 100$ and estimate the AER by the proportion of misclassification in $\mathbf{Z}_1,\ldots,\mathbf{Z}_{2m}$. The simulation size is 1000. Figure 15.1 shows the AER of three bivariate spherically symmetric distributions, namely normal, Laplace and t with 3 degrees of freedom, as δ varies in $[-2,2]$. As expected, the misclassification error is nearly 0.5 when $\delta = 0$ and it decreases as δ goes away from 0 and the separation between the population increases.

In Fig. 15.2, we present a comparison of the proposed classifier (RDC) (given in (15.5)) with some existing classifiers like Fisher's LDA, support vector machine (SVM) (Vapnik 1982, 1998), maximal depth classifier based on Oja depth (denoted by O-D) and projection depth (denoted by P-D) (Ghosh and Chaudhuri 2005b), as well as minimal rank classifier (RC) as proposed by Makinde (2014). We use

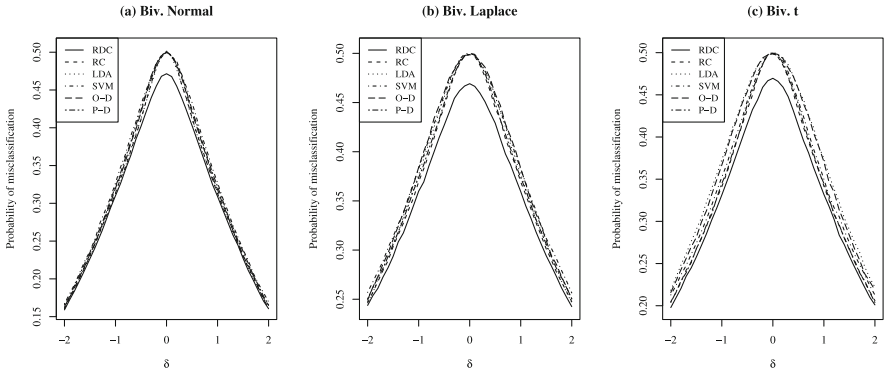

Fig. 15.2 Comparison of several classifiers for some spherically symmetric distributions with a location shift only

the same simulation setting as before. For SVM, we have used *C*-support vector classification with Gaussian RBF-kernel. The default choice of parameters in the R package `kernlab` were used. The cost of constraint C is taken to be 1 with hyperparameter σ is determined automatically by the `sigest` function in the same library and it returns a value between the 0.1 and 0.9 quantile of $\|\mathbf{x} - \mathbf{x}'\|$. As expected we observe that the misclassification errors of all of them are very similar, validating the fact that all of them are equivalent to the Bayes' rule for spherically symmetric distributions with a location shift only.

Next, we investigate the performance of our proposed classifier, RDC, when the populations are not spherically symmetric. To begin with let us consider the classes π_1 and π_2 that have the same elliptically symmetric distribution with the same scale matrix $\boldsymbol{\Sigma}$, but different location vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. In our simulation study, again we consider $d = 2$, $\boldsymbol{\mu}_1 = (0, 0)^\top$ and $\boldsymbol{\mu}_2$ is such that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \delta^2$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Table 15.1 shows the estimated actual error rates of RDC for different values of ρ and δ with three different distributions normal, Laplace and t with 3 degrees of freedom. For comparison, we have also presented the estimated AER for the LDA, which does not depend on ρ. We observe that the error rates for RDC are different for different values of ρ indicating that the proposed classifier is not affine invariant. We know that the spatial rank function is not invariant under general affine transformations and hence any procedure based on spatial ranks is not affine invariant.

Table 15.1 Estimated AER for RDC in the presence of correlation among the variables

Distribution	δ	LDA	RDC			
			$\rho = 0$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.9$
Normal	0.0	0.5000	0.4702	0.4696	0.4704	0.4704
	0.5	0.4059	0.4031	0.4040	0.4031	0.4023
	1.0	0.3117	0.3107	0.3153	0.3155	0.3131
	1.5	0.2290	0.2283	0.2352	0.2345	0.2313
	2.0	0.1602	0.1583	0.1677	0.1674	0.1626
Laplace	0.0	0.5000	0.4703	0.4703	0.4698	0.4699
	0.5	0.4361	0.4278	0.4265	0.4259	0.4259
	1.0	0.3577	0.3589	0.3587	0.3584	0.3568
	1.5	0.2960	0.2571	0.2602	0.2598	0.2570
	2.0	0.2434	0.2004	0.2043	0.2038	0.2008
t with 3 d.f.	0.0	0.5000	0.4699	0.4701	0.4703	0.4702
	0.5	0.4216	0.4137	0.4121	0.4117	0.4111
	1.0	0.3347	0.3320	0.3333	0.3332	0.3311
	1.5	0.2618	0.2571	0.2602	0.2598	0.2570
	2.0	0.2018	0.2004	0.2043	0.2038	0.2008

It can be noted that the spatial rank function can be made affine invariant with the following definition

$$rank_F^*(\mathbf{x}) = E_F\left(\frac{\mathbf{V}^{-1}(\mathbf{x}-\mathbf{X})}{\|\mathbf{V}^{-1}(\mathbf{x}-\mathbf{X})\|}\right) \tag{15.7}$$

where $\mathbf{V}$ is a $d\times d$ matrix such that $\mathbf{V}\mathbf{V}^\top = c\boldsymbol{\Sigma}$ for some constant c. If the covariance of the distribution F exists, we can take $\mathbf{V}$ to be the Cholesky decomposition of the covariance matrix. For the empirical versions, one can estimate $\boldsymbol{\Sigma}$ by minimum covariance determinant (MCD) estimator of Rousseeuw (1984) and then $\mathbf{V}$ by its square root matrix. Note that the Cholesky decomposition of $\boldsymbol{\Sigma}$ (or, its estimate) may not produce an affine invariant rank function but the outlyingness function $r_F^*(\mathbf{x})$ based on $rank_F^*(\mathbf{x})$ will be affine invariant and that will make our classification rule affine invariant. In our numerical examples, we refer to this classification rule as RDA-A. In our later simulation studies, we have also used moment based estimators of $\boldsymbol{\Sigma}$ and affine invariant ranks as proposed in Chakraborty (2001). The corresponding classifiers are denoted as RDA-A0 and AIRDC, respectively.

In Table 15.2, we present a comparison of the proposed affine invariant classifier AIRDC with LDA, SVM, maximal depth classifiers based on Oja depth (O-D) and the projection depth (P-D). The simulation settings are exactly the same as in Fig. 15.2, but only for $\delta = 1$ and 2. Since all the classifiers under comparison except RDC are affine invariant, we have taken $\boldsymbol{\Sigma}$ to be the identity matrix. We observe that AIRDC performs quite well compared to other classifiers and the Bayes risk for

Table 15.2 Error rate comparisons of AIRDC with other classifiers when $\boldsymbol{\Sigma} = \mathbf{I}_2$, $\boldsymbol{\mu}_1 = (0,0)^\top$ and $\boldsymbol{\mu}_2 = (\delta,0)^\top$

		Classifiers							
Distribution	δ	Bayes Risk	LDA	SVM	O-D	P-D	RC	RDC	AIRDC
Normal	1.0	0.3085	0.3148	0.3157	0.3181	0.3213	0.3129	0.3104	0.3096
	2.0	0.1587	0.1612	0.1602	0.1649	0.1660	0.1615	0.1583	0.1582
Laplace	1.0	0.3576	0.3770	0.3814	0.3831	0.3729	0.3693	0.3585	0.3598
	2.0	0.2415	0.2464	0.2573	0.2503	0.2508	0.2475	0.2411	0.2442
t with 3 d.f.	1.0	0.3339	0.3746	0.3505	0.3707	0.3400	0.3418	0.3329	0.3320
	2.0	0.2019	0.2220	0.2137	0.2185	0.2053	0.2060	0.1983	0.1970

normal distribution. For non-normal distributions, it is noticeably better than other classifiers.

If $rank^*_{F_j}(\mathbf{x})$ is an affine invariant spatial rank function based on F_j and $F_{R^*_j}$ is the distribution function of $r^*_{F_j}(\mathbf{X}_j) = \|rank^*_{F_j}(\mathbf{X}_j)\|$ where the d-dimensional random vectors $\mathbf{X}_j$ have distributions F_j for $j = 1, \ldots, K$, we have the following optimal result for AIRDC based on affine invariant rank functions.

Theorem 15.3 *Let F_j be elliptically symmetric distributions with density functions of the form*

$$f_j(\mathbf{x}) = |\boldsymbol{\Sigma}|^{-1/2} h((\mathbf{x}-\boldsymbol{\theta}_j)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\theta}_j))$$

for some strictly decreasing, continuous, non-negative scalar function h. Then the classifier

$$\text{assign } \mathbf{x} \text{ to } \pi_j \text{ if } F_{R^*_j}(r^*_{F_j}(\mathbf{x})) = \min_{1\le i\le K} F_{R^*_i}(r^*_{F_i}(\mathbf{x})) \tag{15.8}$$

is equivalent to the Bayes rule when the prior probabilities $p_1 = \cdots = p_K = 1/K$.

So far we have considered only the populations π_1 and π_2 which are separated by location shift only. In Table 15.3, we present a comparison of error rates of our proposed classifier AIRDC with quadratic discriminant classifier, QDA and RDC when the populations differ by both location and scale. It is well known that *QDA* is the Bayes rule when the two populations are normally distributed and differ in both scale and location. We have taken $\boldsymbol{\Sigma}_1 = \mathbf{I}$, $\boldsymbol{\Sigma}_2 = \sigma^2\mathbf{I}$, $\boldsymbol{\mu}_1 = (0,0)^\top$ and $\boldsymbol{\mu}_2 = (\delta,0)^\top$. We observe that the performance of AIRDC is not very good compared to QDA.

Table 15.3 Error rate comparisons of AIRDC with RDC and QDA when $\boldsymbol{\Sigma}_1 = \mathbf{I}_2$, $\boldsymbol{\Sigma}_2 = \sigma^2\mathbf{I}_2$, $\boldsymbol{\mu}_1 = (0,0)^\top$ and $\boldsymbol{\mu}_2 = (\delta,0)^\top$

Distribution	Classifiers	δ	Average misclassification error				
			$\sigma = 0.20$	$\sigma = 0.50$	$\sigma = 1.0$	$\sigma = 2.0$	$\sigma = 5.0$
Normal	AIRDC	1	0.2943	0.2892	0.3105	0.4179	0.4785
	RDC		0.2892	0.2921	0.3101	0.3578	0.4345
	QDA		0.0556	0.1889	0.3136	0.2448	0.0813
	AIRDC	2	0.0695	0.0987	0.1589	0.2895	0.4496
	RDC		0.0943	0.1258	0.1589	0.2102	0.3167
	QDA		0.0184	0.0751	0.1614	0.1891	0.0784
Laplace	AIRDC	1	0.3594	0.3522	0.3614	0.4362	0.4789
	RDC		0.3616	0.3521	0.3569	0.4377	0.4793
	QDA		0.1045	0.2714	0.3766	0.3109	0.1436
	AIRDC	2	0.2019	0.2038	0.2425	0.3523	0.4586
	RDC		0.2028	0.2073	0.2423	0.3537	0.4586
	QDA		0.0564	0.1613	0.2471	0.2709	0.1379
t with 3 d.f.	AIRDC	1	0.3187	0.3153	0.3376	0.4198	0.4782
	RDC		0.3197	0.3152	0.3320	0.4236	0.4777
	QDA		0.1123	0.2685	0.3716	0.3295	0.1611
	AIRDC	2	0.1408	0.1577	0.2031	0.3144	0.4509
	RDC		0.1449	0.1560	0.2010	0.3143	0.4514
	QDA		0.0510	0.1373	0.2225	0.2692	0.1547

15.4 Numerical Examples with Real Data

We analyse seven benchmark data sets to illustrate the performances of our methods (RDC, AIRDC and RDA-A). These datasets include iris data (Fisher 1936), Pima Indians diabetes (PID) data (owned by the National Institute of Diabetes and Digestive and Kidney Diseases), banknote data (Lohweg 2013), biomedical data (Cox et al. 1982), yeast data (Nakai 1991), cloud data (Miller et al. 1979), seed data (Charytanowicz et al. 2012) and Haberman data (see Haberman 1976). Except for iris data, all data sets are taken from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.html) and StatLib Archive (http://lib.stat.cmu.edu/datasets/). In Table 15.4, we present the analysis of some real data sets with training sample size of each class n and test sample size m (unless mentioned otherwise). Here d is the dimension of the data and k is the number of classes considered. For clarity in our data analysis, we denote RDA-A with MCD estimate of covariance by RDA-A and RDA-A with moment estimate of covariance by RDA-A0. For computing MCD estimate of covariance using the FAST-MCD algorithm (Rousseeuw and Van Driessen 1999) via R package `robustbase`, we set $\alpha = 0.90$ for small training sample sizes (iris data, seed data and biomedical data) and $\alpha = 0.70$ for large training sample sizes. For iris data, RDA-A, RDA-A0, AIRDC and AIRC have the same misclassification error as QDA while RC and RDC have

Table 15.4 Estimated error rates of the classifiers in the real data examples

					Classifiers								
Dataset	n	m	d	k	LDA	QDA	SVM	RC	RDC	AIRC	RDA A	RDA A0	AIRDC
Iris	30	20	4	3	0.033	0.017	0.067	0.033	0.033	0.017	0.0167	0.017	0.017
Biomedical	50	17	4	2	0.206	0.147	0.176	0.206	0.177	0.147	0.147	0.118	0.147
PID	100	100	4	2	0.260	0.280	0.275	0.360	0.350	0.285	0.260	0.265	0.280
Cloud	40	14	4	2	0.571	0.429	0.393	0.464	0.464	0.357	0.286	0.393	0.357
Banknote	100	100	4	2	0.020	0.020	0.005	0.205	0.195	0.020	0.130	0.015	0.015
Seed	50	20	6	3	0.117	0.133	0.167	0.250	0.250	0.117	0.183	0.167	0.117
Haberman	$n_1 = 150$, $n_2 = 50$	$m_1 = 50$, $m_2 = 30$	2	2	0.315	0.258	0.27	0.430	0.400	0.390	0.348	0.392	0.392
Yeast	100	100	4	4	0.495	0.360	0.74	0.550	0.530	0.388	0.448	0.375	0.353

the same misclassification error as LDA. For biomedical data, RDA-A0 has the least misclassification error. RDA-A and AIRDC perform similar to QDA. For Pima Indian diabetes data, RDA-A, RDA-A0 and LDA appear to perform best while QDA and AIRDC perform well. For cloud data, RDA-A outperforms others while AIRC, AIRDC and RDA-A0 outperform QDA and LDA. AIRC and RDA-A compete favourably with all other classifiers for banknote authentication data, while misclassification error is least in RDA-A0 but high in RDA-A, RC and RDC. For seed data, both AIRDC, AIRC and LDA outperform others. For Haberman data, QDA has the least misclassification error while RDA-A and LDA perform well. AIRDC, QDA, RDA-A0 and AIRC perform best among other classifiers with yeast data. In general, AIRC, AIRDC, RDA-A and RDA-A0 perform well and compete favourably with QDA while RC and RDC compete favourably with LDA. For comparison, we have also included error rates of the SVM classifier with the settings as described earlier. We have not included OD and PD classifiers due to high dimensionality of the data sets.

15.5 Concluding Remarks

The classifier proposed in (15.5) based on the training samples does not depend on any distributional assumptions or does not require any estimation of model parameters. That gives a nonparametric flavour to our classifier RDC. It is also computationally simple and can be applied to very high dimensional data as well. In principle, the proposed methods can be applied to high dimensional situation when dimension $d > n$, the sample size, however, the properties and the behaviour of the classifier in those circumstances require further investigation. We have noted in our real data examples that the proposed classifiers are quite competitive with similar classifiers but computationally much easier than maximal depth based classifiers (except may be for spatial depth). Due to computational complexity of the depth functions for dimensions $d \geq 3$, it is nearly impossible to implement those classifiers for large dimensions unless we have a very small sample size. However, the classifiers proposed here can be computed in any dimension quite easily. In future work, we propose to investigate the usefulness of these classifiers when dimension $d > n$.

References

Chakraborty, B.: On affine equivariant multivariate quantiles. Ann. Inst. Stat. Math. **53**, 380–403 (2001)

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P.A., Lukasik, S. Zak, S.: UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science (2012). http://mlr.cs.umass.edu/ml/datasets/seeds

Christmann, A., Rousseeuw, P.: Measuring overlap in binary regression. Comput. Stat. Data Anal. **37**, 65–75 (2001)
Christmann, A., Fischer, P., Joachims, T.: Comparison between various regression depth methods and the support vector machine to approximate the minimum number of misclassifications. Comput. Stat. **17**, 273–287 (2002)
Cox, L.H., Johnson, M.M., Kafadar, K.: Exposition of statistical graphics technology. ASA Proceedings of the Statistical Computation Section, pp. 55–56 (1982)
Cui, X., Lin, L., Yang, G.R.: An extended projection data depth and its applications to discrimination. Commun. Stat.-Theory Methods **37**, 2276–2290 (2008)
Durrett, R.: Probability: Theory and Examples, 4th edn. Cambridge University Press, Boston (2010)
Dutta, S., Ghosh, A.K.: On robust classification using projection depth. Ann. Inst. Stat. Math. **64**, 657–676 (2012a)
Dutta, S. and Ghosh, A. K.: On classification based on L_p depth with an adaptive choice of p. Technical Report No. R5/2011, Statistics and Mathematics Unit. Indian Statistical Institute, Kolkata, India (2012b)
Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**, 179–188 (1936)
Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, New York (1990)
Ghosh, A.K. Chaudhuri, P.: On data depth and distribution-free discriminant analysis using separating surfaces. Bernoulli **11**, 1–27 (2005a)
Ghosh, A.K., Chaudhuri, P.: On maximum depth and related classifiers. Scand. J. Stat. **32**, 327–350 (2005b)
Guha, P.: On scale-scale curves for multivariate data based on rank regions. Ph.D. thesis, University of Birmingham (2012)
Guha, P., Chakraborty, B.: On scale-scale plot for comparing multivariate distributions. Submitted (2013)
Haberman, S.J.: Generalized residuals for log-linear models. Proceedings of the 9th International Biometrics Conference, pp. 104–122. Boston (1976)
Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, New York (2001)
Jörnsten, R.: Clustering and Classification based on the L_1 data depth. J. Multivar. Anal. **90**, 67–89 (2004)
Koltchinskii, V.I.: M-estimation, convexity and quantiles. Ann. Stat. **25**, 435–477 (1997)
Lachenbruch, P.A., Mickey, M.R.: Estimation of error rates in discriminant analysis. Technometrics **10**, 1–11 (1968)
Lange, T., Mosler, K., Mozharovskyi, P.: Fast nonparametric classification based on data depth. Stat. Pap. **55**, 49–69 (2014)
Li, J., Cuesta-Albertos, J. A. and Liu, R. Y.: DD-classifier: Nonparametric classification procedure based on DD-plot. J. Am. Stat. Assoc. **107**, 737–753 (2012)
Liu, R.Y.: Control charts for multivariate processes. J. Am. Stat. Assoc. **90**, 1380–1387 (1995)
Liu, R.Y., Parelius, J.M., Singh, K.: Multivariate analysis by data depth: Descriptive statistics, graphics and inference. Ann. Stat. **27**, 783–858 (1999)
Liu, R.Y., Singh, K.: A quality index based on multivariate data depth and multivariate rank tests. J. Am. Stat. Assoc. **88**, 252–260 (1993)
Lohweg, V.: UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science (2013). http://mlr.cs.umass.edu/ml/datasets/banknote+authentication
Makinde, O.S.: On some classification methods for high dimensional and functional data. PhD Thesis, University of Birmingham (2014)
Marden, J.I.: Bivariate QQ-Plots and spider web plots. Stat. Sin. **8**, 813–826 (1998)
McLachlan, G.J.: Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York (1992)

Miller, A.J., Shaw, D.E., Veitch, L.G., Smith, E.J.: Analyzing the results of a cloud-seeding experiment in Tasmania. Commun. Stat.-Theory Methods **8**, 1017–1047 (1979)
Möttönen, J., Oja, H.: Multivariate spatial sign and rank methods. J. Nonparametric Stat. **5**, 201–213 (1995)
Möttönen, J., Oja, H., Tienari, J.: On the efficiency of multivariate spatial sign and rank tests. Ann. Stat. **25**, 542–552 (1997)
Nakai, K.: UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science (1991). http://mlr.cs.umass.edu/ml/datasets/Yeast
Paindaveine, D., Van Bever, G.: Nonparametrically consistent depth-based classifiers. Bernoulli **21**, 62–82 (2015)
Rousseeuw, P.J.: Least median of squares regression. J. Am. Stat. Assoc. **79**, 871–880 (1984)
Rousseeuw, P.J., Hubert, M.: Regression depth. J. Am. Stat. Assoc. **94**, 388–402 (1999)
Rousseeuw, P.J., Van Driessen, K.: A fast algorithm for the minimum covariance determinant estimator. Technometrics **41**, 212–223 (1999)
Serfling, R.: A depth function and a scale curve based on spatial quantiles. In: Dodge, Y. (ed.) Statistical Data Analysis Based on the L_1-Norm and Related Methods, pp. 25–28. Birkhaeuser, Boston (2002)
Serfling, R.: Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardization. J. Nonparametric Stat. **22**, 915–936 (2010)
Vapnik, V.N.: Estimation of Dependences Based on Empirical Data. Addendum 1. Springer, New York (1982)
Vapnik V.N.: Statistical Learning Theory. Wiley, New York (1998)
Vardi, Y., Zhang, C.H.: The multivariate L_1-median and associated data depth. Proc. Natl. Acad. Sci. **97**, 1423–1426 (2000)

Chapter 16
Robust Change Detection in the Dependence Structure of Multivariate Time Series

Daniel Vogel and Roland Fried

Abstract A robust change-point test based on the spatial sign covariance matrix is proposed. A major advantage of the test is its computational simplicity, making it particularly appealing for robust, high-dimensional data analysis. We derive the asymptotic distribution of the test statistic for stationary sequences, which we allow to be near-epoch dependent in probability (P NED) with respect to an α-mixing process. Contrary to the usual L_2 near-epoch dependence, this short-range dependence condition requires no moment assumptions, and includes arbitrarily heavy-tailed processes. Further, we give a short review of the spatial sign covariance matrix and compare our test to a similar one based on the sample covariance matrix in a simulation study.

Keywords GARCH • Near epoch dependence • Oja sign covariance matrix • Orthogonal invariance • Spatial sign covariance matrix • Tyler matrix

16.1 Introduction

The problem of detecting changes in the parameters of time series has been an area of intensive research for several decades. However, detecting breaks in the dependence structure of multivariate time series has just recently been addressed in the literature (e.g., Galeano and Peña 2007; Aue et al. 2009; Wied et al. 2012; Dehling et al. 2012; Quessy et al. 2013; Galeano and Wied 2014; Wied et al. 2014; Bücher et al. 2014). The majority of these articles study variants of the general CUSUM-type test statistic, being of the form

$$\hat{T}_n = \max_{1 \le k \le n} \frac{k}{\hat{\sigma}_{\theta,n}\sqrt{n}} |\hat{\theta}_k - \hat{\theta}_n|, \tag{16.1}$$

D. Vogel (✉)
Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB24 3UE, UK
e-mail: daniel.vogel@abdn.ac.uk

R. Fried
Fakultät Statistik, Technische Universität Dortmund, D-44221 Dortmund, Germany
e-mail: fried@statistik.tu-dortmund.de

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_16

where $\hat{\theta}_k$, $1 \leq k \leq n$, is an estimator based on the first k observations for the parameter of interest θ, and $\hat{\sigma}^2_{\theta,n}$ is an appropriate variance estimator. In this formulation, the parameter θ is univariate, but the basic principle of $\hat{\theta}_k$ being compared to $\hat{\theta}_n$ for all $1 \leq k \leq n$ and the appropriate scaling with respect to k and n also apply to parameter vectors. In Aue et al. (2009), for instance, the parameter of interest is the covariance matrix, and in Galeano and Wied (2014) it is the correlation matrix. It is well known that the respective normal maximum likelihood estimators possess certain optimality properties under normality, but are rather inefficient under heavy tails. The latter is a serious drawback in practice. The tests are mainly intended for applications to financial time series, which are, even on a log-scale, known to display heavy tails. The tests of Aue et al. (2009) and Galeano and Wied (2014) require finite fourth moments, and it is a matter of debate whether this may be assumed for, e.g., log-returns of stock prices. Regardless of the answer to this specific question, the tests are used in data situations where they are far from optimal. Several researchers have addressed this issue by proposing change-point tests based on alternative estimators. For instance, Dehling et al. (2012) and Quessy et al. (2013) consider Kendall's τ, and Wied et al. (2014) and Kojadinovic et al. (2015) propose test statistics related to Spearman's ρ. However, these robustifications come at a higher computational cost, making them less feasible for change detection in high dimensions or an on-line setting. In the present paper, we propose a change-point test with very good robustness properties, which requires no moment assumptions at all, comes at no higher computational cost as, e.g., the sample covariance based test by Aue et al. (2009), and also has a surprisingly high efficiency under normality.

The paper is organized as follows. In Sect. 16.2, we state the problem, define the test statistic, and derive its asymptotic null distribution. Our test statistic is based on the spatial sign covariance matrix, which we review in Sect. 16.3. Then Sect. 16.4 contains numerical simulation results, and we close with a discussion and an outlook in Sect. 16.5.

Throughout, we use $\|\cdot\|_r$ to denote the r norm in $\mathbb{R}^p$ for any $1 \leq r \leq \infty$, and write $\|\cdot\|$ short for $\|\cdot\|_2$. Thus for a random vector $\mathbf{X}$, $\|\mathbf{X}\|_2$ is a univariate random variable. We use $\{E\|\mathbf{X}\|_r^r\}^{1/r}$ to denote the L_r norm of $\mathbf{X}$. All random variables are defined on a common probability space $(\Omega, \mathscr{F}, P)$. Finally, $\sigma(\mathbf{X}_i \mid i \in I)$ denotes the σ-field generated by $\mathbf{X}_i$, $i \in I$, where I is some arbitrary index set.

16.2 Statement of Problem and Main Result

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a p-variate random sample with $p \geq 2$. The spatial median of the (distribution of the) random vector $\mathbf{X}$ is defined as

$$\boldsymbol{\mu}(\mathbf{X}) = \underset{\boldsymbol{\mu} \in \mathbb{R}^p}{\arg\min}\, E\left(\|\mathbf{X} - \boldsymbol{\mu}\| - \|\mathbf{X}\|\right). \tag{16.2}$$

The spatial median always exists and, if the distribution of $\mathbf{X}$ is not concentrated on a straight line, it is unique. Throughout the paper, we assume the data to be centered by the spatial median, i.e, $\boldsymbol{\mu}(\mathbf{X}_1) = \ldots = \boldsymbol{\mu}(\mathbf{X}_n) = \mathbf{0}$. We completely exclude the location estimation problem and assume the spatial median of the data to be known. The feasibility of this simplifying assumption and the question of how to extend our results to the unknown-location setting will be discussed in Sect. 16.5.

We want to detect changes in the dependence structure of the $\mathbf{X}_k$. By dependence structure we mean the covariance up to scale. Thus, if we denote $\mathbf{Cov}(\mathbf{X})$ by $\boldsymbol{\Sigma}(\mathbf{X})$, and let $\boldsymbol{\Sigma}_0(\mathbf{X}) = \mathrm{tr}(\boldsymbol{\Sigma}(\mathbf{X}))^{-1}\boldsymbol{\Sigma}(\mathbf{X})$, we want to test the hypothesis

$$H_0 : \boldsymbol{\Sigma}_0(\mathbf{X}_1) = \boldsymbol{\Sigma}_0(\mathbf{X}_2) = \ldots = \boldsymbol{\Sigma}_0(\mathbf{X}_n) \tag{16.3}$$

against the alternative that any of these equalities is violated. In particular, we aim at power against one-jump alternatives, i.e.,

$$H_A : \exists\, 1 \leq k < n : \boldsymbol{\Sigma}_0(\mathbf{X}_1) = \ldots = \boldsymbol{\Sigma}_0(\mathbf{X}_k) \neq \boldsymbol{\Sigma}_0(\mathbf{X}_{k+1}) = \ldots = \boldsymbol{\Sigma}(\mathbf{X}_n).$$

We call

$$\mathbf{s}(\mathbf{x}) = \begin{cases} \mathbf{x}/\|\mathbf{x}\| & \text{if } \mathbf{x} \neq \mathbf{0}, \\ \mathbf{0} & \text{otherwise,} \end{cases}$$

the *spatial sign* of $\mathbf{x} \in \mathbb{R}^p$. Let $\mathrm{vec}(\mathbf{A})$ be the p^2 vector obtained by stacking the columns of $\mathbf{A} \in \mathbb{R}^{p\times p}$ from left to right underneath each other, and $\mathrm{vech}(\mathbf{A})$ be the $p(p+1)/2$ vector obtained from $\mathrm{vec}(\mathbf{A})$ by removing all subdiagonal elements of $\mathbf{A}$. We define the function $\mathbf{v}$ as

$$\mathbf{v} : \mathbb{R}^p \to \mathbb{R}^m : \mathbf{x} \mapsto \tilde{\mathbf{I}}_m \,\mathrm{vech}\left\{\mathbf{s}(\mathbf{x})\mathbf{s}(\mathbf{x})^\top\right\}, \tag{16.4}$$

where $m = p(p+1)/2 - 1$, and $\tilde{\mathbf{I}}_m$ denotes the $m \times (m+1)$ matrix obtained from the $(m+1)$-dimensional identity matrix $\mathbf{I}_{m+1}$ by removing the last line. Let further

$$\boldsymbol{\Phi}_k = \frac{1}{\sqrt{n}} \left\{ \sum_{j=1}^{k} \mathbf{v}(\mathbf{X}_j) - \frac{k}{n} \sum_{j=1}^{n} \mathbf{v}(\mathbf{X}_j) \right\} \tag{16.5}$$

and

$$\hat{\boldsymbol{\Omega}}_{n,0} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{v}(\mathbf{X}_i)\mathbf{v}(\mathbf{X}_i)^\top. \tag{16.6}$$

Our change-point test statistic for testing H_0 against H_A is then

$$\hat{T}_{S,0} = \max_{1 \le k \le n} \boldsymbol{\Phi}_k^\top \hat{\boldsymbol{\Omega}}_{n,0}^{-1} \boldsymbol{\Phi}_k.$$

The following theorem gives the asymptotic null distribution of $\hat{T}_{S,0}$ for independent sequences.

Theorem 16.1 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n, \ldots$ be p-dimensional, independent, and identically distributed observations with distribution F, which is not concentrated on a $(p-1)$-dimensional hyperplane. Then, as $n \to \infty$,*

$$\hat{T}_{S,0} \xrightarrow{d} T(m) = \sup_{0 \le t \le 1} \sum_{l=1}^{m} B_l^2(t), \tag{16.7}$$

where B_l, $1 \le l \le m$, are independent standard Brownian motions on $[0, 1]$.

We give no proof of Theorem 16.1 since it is a special case of Theorem 16.2 below. The distribution of $T(m)$ can easily be simulated, and has also been tabulated, e.g. by Aue et al. (2009). The authors also state that

$$(m/8)^{-1/2}(T(m) - m/4) \xrightarrow{\text{d}} N(0,1) \qquad \text{as } m \to \infty,$$

but note that this convergence is too slow to provide sensible critical values for practical purposes.

Remark 16.1 Noteworthy about Theorem 16.1 is that it poses basically no restriction on the population distribution F, in particular no moment condition. The latter may appear superfluous, since the formulation of the hypothesis H_0, cf. (16.3), implicitly requires the existence of finite second moments. However, there are alternative ways of modeling multivariate dependencies and multivariate scatter, which are not based on the covariance matrix and which also extend beyond finite second moments. The test statistic $\hat{T}_{S,0}$ may be put to good use in such situations as well. We will discuss in some detail in Sect. 16.3 how this test can be employed to test for constancy of the shape of an elliptical distribution, regardless of the existence of any moments.

Remark 16.2 The requirement for F to be not concentrated on an affine-linear subspace is also not necessary, but a mere convenience assumption. If F was confined to a, say, $k < p$ dimensional hyperplane, then the sum in (16.7) would range from 1 to $k(k+1)/2-1$ and $\hat{\boldsymbol{\Omega}}_{n,0}^{-1}$ would have to be replaced by a generalized inverse. It should also be noted that, if the spatial median $\boldsymbol{\mu}$ was to be estimated, some regularity condition, similar to condition (ii) of Theorem 16.2, prohibiting a strong probability mass concentration at $\boldsymbol{\mu}$, would be necessary.

The assumption of independent observations will be relaxed below. Before doing so, we want to motivate the construction of the test statistic $\hat{T}_{S,0}$ and explain that it is indeed a multivariate version of (16.1). For that purpose, we briefly recall the test statistic of Aue et al. (2009). Following the construction principle (16.1), Aue et al. (2009) consider

$$\hat{T}_{\boldsymbol{\Sigma}} = \max_{1\le k\le n} \boldsymbol{\Psi}_k^\top \hat{\boldsymbol{\Gamma}}_n^{-1} \boldsymbol{\Psi}_k, \quad \text{where } \boldsymbol{\Psi}_k = \operatorname{vech}\left\{ \frac{k}{\sqrt{n}} (\hat{\boldsymbol{\Sigma}}_k - \hat{\boldsymbol{\Sigma}}_n) \right\}, \quad 1 \le k \le n, \tag{16.8}$$

for testing the constancy of the covariance matrix $\boldsymbol{\Sigma}$. Here, $\hat{\boldsymbol{\Sigma}}_k$ denotes the sample covariance matrix computed from the first k observations (but with the location estimated from the whole sample), and $\hat{\boldsymbol{\Gamma}}_n$ is a suitable long-run variance estimator for the statistic $\operatorname{vech} \hat{\boldsymbol{\Sigma}}_n$. Thus the multivariate formulation of (16.1) is in the same vein as the classical univariate t-test generalizes to Hotelling's T^2. Ignoring the location estimation, $\boldsymbol{\Psi}_k$ can be written as

$$\boldsymbol{\Psi}_k = \frac{1}{\sqrt{n}} \left\{ \sum_{j=1}^{k} \operatorname{vech}(\mathbf{X}_j \mathbf{X}_j^\top) - \frac{k}{n} \sum_{j=1}^{n} \operatorname{vech}(\mathbf{X}_j \mathbf{X}_j^\top) \right\}.$$

Our test statistic is in analogy to Aue et al. (2009), except that the function $\mathbf{x} \mapsto \operatorname{vech}(\mathbf{x}\mathbf{x}^\top)$ is replaced by $\mathbf{v}(\cdot)$, cf. (16.4). Why one would consider the function $\mathbf{v}$ at this point will be discussed in Sect. 16.3. That this choice leads to a very useful change-point test is demonstrated in Sect. 16.4.

In the remainder of this section, we extend Theorem 16.1 to dependent observations. When proving results for dependent sequences, one requires technical short-range dependence conditions that capture the degree of dependence. By now, a multitude of versions and variants of such conditions exist, which differ in the broadness of the processes covered, the strength of the results obtainable and in the effort necessary to prove limit theorems and to check whether specific models satisfy the conditions. Consequently, every author has certain personal preferences. In the present situation, any set of conditions is fine as long as convergence of the partial sum process of $(\mathbf{v}(\mathbf{X}_n))_{n\in\mathbb{N}}$ to a multivariate Brownian motion can be shown and a consistent long-run variance estimator for

$$\boldsymbol{\Omega} = \lim_{n\to\infty} \boldsymbol{\Omega}_n \quad \text{with} \quad \boldsymbol{\Omega}_n = \mathbf{C}\mathrm{ov}\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{v}(\mathbf{X}_i) \right\}$$

can be provided. Pursuing further the analogy to Aue et al. (2009), their Assumption 2.1 is one possibility. However, one decisive advantage of Theorem 16.1 is that no moment conditions whatsoever are required, and we would like to maintain this moment-freeness also in the dependent setting and avoid any moments conditions implicitly hidden in the short-range dependence condition. We therefore use the P NED condition (near-epoch dependence in probability), proposed by Dehling et al.

(2012), which was introduced for exactly this purpose: to formulate limit theorems for dependent and heavy-tailed processes.

Definition 16.1 The p-dimensional process $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ is called *near-epoch dependent in probability* or short *P near-epoch dependent* (P NED) on the q-dimensional process $(\mathbf{Z}_n)_{n\in\mathbb{Z}}$ if there is a sequence of approximating constants $(a_k)_{k\in\mathbb{N}}$ with $a_k \to 0$ as $k \to \infty$, a sequence of functions $\mathbf{f}_k : \mathbb{R}^{q\times(2k+1)} \to \mathbb{R}^p$, $k \in \mathbb{N}$, and a non-increasing function $\phi : (0,\infty) \to (0,\infty)$ such that

$$P\left(\|\mathbf{X}_0 - \mathbf{f}_k(\mathbf{Z}_{-k},\ldots,\mathbf{Z}_k)\| > \varepsilon\right) \;\le\; a_k\phi(\varepsilon)$$

for all $k \in \mathbb{N}$ and $\varepsilon > 0$.

Here, it is assumed that $((\mathbf{X}_n,\mathbf{Z}_n))_{n\in\mathbb{Z}}$ is a strongly stationary process, and that the observed data is the positive branch $(\mathbf{X}_n)_{n\in\mathbb{N}}$ of the doubly infinite sequence $(\mathbf{X}_n)_{n\in\mathbb{Z}}$. The underlying process $(\mathbf{Z}_n)_{n\in\mathbb{Z}}$ itself can be an i.i.d. process, an m-dependent process (i.e., any two observations being apart more than lag m are independent), or, more generally, a mixing process. For brevity, we focus here solely on the most general of the available mixing conditions, α-mixing, but remark that the result can be stated likewise for φ-mixing.

Definition 16.2 The q-dimensional, stationary process $(\mathbf{Z}_n)_{n\in\mathbb{N}}$ is called *α-mixing* or *strongly mixing* with coefficients $(\alpha_k)_{k\in\mathbb{N}}$ if

$$\alpha_k = \sup_{A\in\mathscr{F}^0_{-\infty}} \sup_{B\in\mathscr{F}^\infty_k} |P(A\cap B) - P(A)P(B)| \;\to 0 \qquad \text{as } k\to\infty,$$

where $\mathscr{F}^0_{-\infty} = \sigma(\ldots,\mathbf{X}_{-1},\mathbf{X}_0)$ and $\mathscr{F}^\infty_k = \sigma(\mathbf{X}_k,\mathbf{X}_{k+1},\ldots)$.

Further, we require a long-run variance estimator that accounts for the serial dependence of the observations. The estimator $\hat{\boldsymbol{\Omega}}_{n,0}$, defined in (16.6), is not consistent for $\boldsymbol{\Omega}$ under dependence. Let

$$\hat{\boldsymbol{\Omega}}_{n,b_n} = \frac{1}{n}\sum_{k=\lfloor -b_n\rfloor}^{\lceil b_n\rceil}\left(1-\frac{|k|}{b_n+1}\right)\sum_{j=1}^{n-|k|}\mathbf{v}(\mathbf{X}_j)\mathbf{v}(\mathbf{X}_{j+k}), \tag{16.9}$$

where $b_n \ge 0$ is a bandwidth parameter that increases with n. Then $\hat{\boldsymbol{\Omega}}_{n,0}$ is a special case of $\hat{\boldsymbol{\Omega}}_{n,b_n}$ obtained for $b_n = 0$. We are now ready to formulate the main result of the paper.

Theorem 16.2 *Let $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ be a p-dimensional, stationary process with marginal distribution F, such that*

(i) F is not concentrated on a lower dimensional hyperplane and
(ii) $P(\|\mathbf{X}_0\| \le \epsilon) = O(\epsilon^2)$ as $\epsilon \to 0$.
Let $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ be P NED with function ϕ and approximating constants $(a_k)_{k\ge1}$ on an α-mixing process with mixing coefficients $(\alpha_k)_{k\ge1}$ satisfying

(iii) $a_k\phi(k^{-1-\delta}) = O(k^{-1-\delta})$ *and*
(iv) $\alpha_k = O(k^{-1-\delta})$

for some $\delta > 0$ *as* $k \to \infty$. *Let furthermore* $(b_n)_{n\in\mathbb{N}}$ *be a non-decreasing sequence such that* $b_n \to \infty$ *and* $b_n = o(n^{1/2})$ *as* $n \to \infty$. *Then*

$$\hat{T}_{S,b_n} = \max_{1\le k\le n} \boldsymbol{\Phi}_k^\top \hat{\boldsymbol{\Omega}}_{n,b_n}^{-1} \boldsymbol{\Phi}_k \xrightarrow{d} T(m),$$

where $\boldsymbol{\Phi}_k$ *is defined in (16.5), and* $T(m)$ *is as in Theorem 16.1.*

Again, for clarity of exposition, we have in the utmost generality. For instance, a large class of kernel functions is possible in (16.9) instead of the Bartlett kernel $(1-|x|)\mathbb{1}_{[-1,1]}(x)$, cf. e.g. de Jong and Davidson (2000, Assumption 1). As compared to Theorem 16.1, we have the additional regularity condition (ii) on the population distribution F, which requires F not to be too strongly concentrated at the origin. This condition is necessary for the near-epoch dependence of $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ to carry over to $(\mathbf{v}(\mathbf{X}_n))_{n\in\mathbb{Z}}$. It is a very mild condition. For instance, if F has a p-dimensional Lebesgue density, continuous at $\mathbf{0}$, then $P(\|\mathbf{X}_0\| \le \epsilon) = O(\epsilon^p)$. The essential prerequisites for proving Theorem 16.2 are the following two lemmas.

Lemma 16.1 *Let* $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ *be a* p*-dimensional, stationary process with marginal distribution* F*, which is* P *NED with function* ϕ *and approximating constants* $(a_k)_{k\ge1}$ *on a stationary process* $(\mathbf{Z}_n)_{n\in\mathbb{Z}}$ *such that there exists a sequence of real numbers* $(s_k)_{k\in\mathbb{N}}$ *with* $a_k\phi(s_k) = O(s_k)$ *as* $k \to \infty$. *Let further* $\mathbf{h} : \mathbb{R}^p \to \mathbb{R}^d$ *be bounded and*

$$E\left\{\sup_{\{\mathbf{x}:\|\mathbf{x}-\mathbf{x}_0\|\le\epsilon\}} \|\mathbf{h}(\mathbf{x}) - \mathbf{h}(\mathbf{X}_0)\|^2\right\} \le L\epsilon \tag{16.10}$$

for all $\epsilon > 0$ *and some constant* $L > 0$ *independent of* ϵ. *Then* $(\mathbf{h}(\mathbf{X}_n))_{n\in\mathbb{Z}}$ *is* L_2 *NED on* $(\mathbf{Z}_n)_{n\in\mathbb{N}}$ *with coefficients* $(a_{2,k})_{k\in\mathbb{N}}$ *satisfying* $a_{2,k} = O(s_k^{1/2})$ *as* $k \to \infty$.

Remark 16.3

(i) A stationary process $(\mathbf{Y}_n)_{n\in\mathbb{Z}}$ is said to be L_r NED, $r \ge 1$, with coefficients $(a_{r,k})_{k\in\mathbb{N}}$ on the stationary process $(\mathbf{Z}_n)_{n\in\mathbb{Z}}$ if

$$a_{r,k} = \{E\|\mathbf{X}_0 - E(\mathbf{X}_0|\sigma(\mathbf{Z}_{-k},\ldots,\mathbf{Z}_k))\|_r^r\}^{1/r}$$

converges to zero for $k \to \infty$.
(ii) If (16.10) is fulfilled, then $\mathbf{h}$ is said to satisfy the *variation condition* with respect to F.

Proof (Lemma 16.1) The P NED assumption of $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ with respect to $(\mathbf{Z}_n)_{n\in\mathbb{Z}}$ states that

$$P(\|\mathbf{X}_0 - \mathbf{f}_k(\mathbf{Z}_{-k},\ldots,\mathbf{Z}_k)\| > s_k) \le a_k\phi(s_k) \le C_1 s_k \tag{16.11}$$

for all $k \in \mathbb{N}$, where $C_1 > 0$ is a constant not depending on k. We abbreviate $\mathbf{f}_k(\mathbf{Z}_{-k}, \ldots, \mathbf{Z}_k)$ by $\tilde{\mathbf{X}}_k$. Since the conditional expectation minimizes the L_2 norm, we have

$$\begin{aligned} a_{2,k}^2 &= E\|\mathbf{h}(\mathbf{X}_0) - E[\mathbf{h}(\mathbf{X}_0)|\sigma(\mathbf{Z}_{-k}, \ldots, \mathbf{Z}_k)]\|^2 \\ &\leq E\|\mathbf{h}(\mathbf{X}_0) - \mathbf{h} \circ \mathbf{f}_k(\mathbf{Z}_{-k}, \ldots, \mathbf{Z}_k)\|^2 \\ &= E\left\{\|\mathbf{h}(\mathbf{X}_0) - \mathbf{h}(\tilde{\mathbf{X}}_k)\|^2 \mathbb{1}_{\{\|\mathbf{X}_0 - \tilde{\mathbf{X}}_k\| \leq s_k\}}\right\} \\ &\quad + E\left\{\|\mathbf{h}(\mathbf{X}_0) - \mathbf{h}(\tilde{\mathbf{X}}_k)\|^2 \mathbb{1}_{\{\|\mathbf{X}_0 - \tilde{\mathbf{X}}_k\| > s_k\}}\right\} \\ &\leq L s_k + C_2 P(\|\mathbf{X}_0 - \tilde{\mathbf{X}}_k\| > s_k) \leq L s_k + C_2 C_1 s_k = O(s_k). \end{aligned}$$

The upper bound on the first term is due to (16.10), the upper bound on the second is due to the boundedness of $\mathbf{h}$ and (16.11). This completes the proof. □

Lemma 16.1 provides the link from the P near-epoch dependence of $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ to the L_2 near-epoch dependence of $(\mathbf{v}(\mathbf{X}_n))_{n\in\mathbb{Z}}$. The next lemma verifies that $\mathbf{v}$, cf. (16.4), satisfies the assumptions of Lemma 16.1 under the conditions of Theorem 16.2.

Lemma 16.2 *Let $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ be as in Theorem 16.2. Then $\tilde{\mathbf{v}} : \mathbb{R}^p \to \mathbb{R}^{p^2} : \mathbf{x} \mapsto \mathrm{vec}\{\mathbf{s}(\mathbf{x})\mathbf{s}(\mathbf{x})^\top\}$ satisfies (16.10).*

Proof (Lemma 16.2) We have

$$\begin{aligned} &E\left[\sup_{\{\mathbf{x}:\|\mathbf{x}-\mathbf{X}_0\|\leq\epsilon\}} \|\tilde{\mathbf{v}}(\mathbf{x}) - \tilde{\mathbf{v}}(\mathbf{X}_0)\|^2\right] \\ &= E\left[\left\{\sup_{\{\mathbf{x}:\|\mathbf{x}-\mathbf{X}_0\|\leq\epsilon\}} \|\tilde{\mathbf{v}}(\mathbf{x}) - \tilde{\mathbf{v}}(\mathbf{X}_0)\|^2\right\} \mathbb{1}_{\{\|\mathbf{X}_0\|\leq\sqrt{\epsilon}\}}\right] \\ &\quad + E\left[\left\{\sup_{\{\mathbf{x}:\|\mathbf{x}-\mathbf{X}_0\|\leq\epsilon\}} \|\tilde{\mathbf{v}}(\mathbf{x}) - \tilde{\mathbf{v}}(\mathbf{X}_0)\|^2\right\} \mathbb{1}_{\{\|\mathbf{X}_0\|>\sqrt{\epsilon}\}}\right] \end{aligned}$$

The first term is bounded by $2P(\|\mathbf{X}_0\| \leq \sqrt{\epsilon}) = O(\epsilon)$ as $\epsilon \to 0$. We deal with the second term as follows. Let $\epsilon < 1$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ such that $\|\mathbf{x}\| > \sqrt{\epsilon}$ and $\|\mathbf{x} - \mathbf{y}\| < \epsilon$. Then

$$\|\mathbf{s}(\mathbf{x}) - \mathbf{s}(\mathbf{y})\| \leq \left\|\frac{\mathbf{x}}{\|\mathbf{x}\|} - \frac{\mathbf{y}}{\|\mathbf{x}\|}\right\| + \left\|\frac{\mathbf{y}}{\|\mathbf{x}\|} - \frac{\mathbf{y}}{\|\mathbf{y}\|}\right\| \leq \frac{\epsilon}{\sqrt{\epsilon}} + \left(\frac{\|\mathbf{y}\|}{\|\mathbf{x}\|} - 1\right) \leq 2\sqrt{\epsilon}.$$

This is equivalent to $\mathbf{s}(\mathbf{x})\mathbf{s}(\mathbf{y})^\top \geq 1 - 2\epsilon$, and consequently $\|\tilde{\mathbf{v}}(\mathbf{x}) - \tilde{\mathbf{v}}(\mathbf{y})\|^2 = 2 - 2\{\mathbf{s}(\mathbf{x})\mathbf{s}(\mathbf{y})\}^2 \leq 2 - 2(1-2\epsilon)^2 \leq 8\epsilon$. Hence also the second term is $O(\epsilon)$ for $\epsilon \to 0$. The proof is complete. □

We are now ready to state the proof of Theorem 16.2.

Proof (Theorem 16.2) The proof requires two ingredients,

(1) the convergence of the partial sum process of $(\mathbf{v}(\mathbf{X}_k))_{k\in\mathbb{N}}$, specifically

$$\boldsymbol{\Omega}_n^{-1/2}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt\rfloor}\{\mathbf{v}(\mathbf{X}_i)-E[\mathbf{v}(\mathbf{X}_i)]\}\right)_{0\le t\le 1}\xrightarrow{\text{d}}(\mathbf{W}(t))_{0\le t\le 1}.$$

for $n\to\infty$ in the p-dimensional Skorokhod space, where $\mathbf{W}$ denotes a p-dimensional standard Brownian motion on $[0,1]$, and

(2) the consistency of the long-run variance estimator $\hat{\boldsymbol{\Omega}}_{n,b_n}$, specifically

$$\|\operatorname{vec}(\hat{\boldsymbol{\Omega}}_{n,b_n}-\boldsymbol{\Omega}_n)\|\xrightarrow{\text{p}}0\qquad\text{as } n\to\infty.$$

From these two results, the claim follows by standard arguments, see, e.g., the proof of Theorem 2.1 in Aue et al. (2009). Part (1) follows with Corollary 4.2 of Wooldridge and White (1988), part (2) with Theorem 1 of de Jong and Davidson (2000). Both propositions are formulated for L_2 NED sequences with respect to an α-mixing process with coefficients satisfying $a_{2,k}=O(k^{-1/2-\delta})$ and $\alpha_k=O(k^{-1-\delta})$ as $k\to\infty$ for some $\delta>0$. Lemmas 16.1 and 16.2 together ensure that these assumptions are met. Assumption D.3 of Wooldridge and White (1988) follows since $\boldsymbol{\Omega}_n\to\boldsymbol{\Omega}$, which can, e.g., be deduced from Ibragimov (Theorem 2.3 1962). □

16.3 The Spatial Sign Covariance Matrix

For a p-dimensional random vector $\mathbf{X}$ with distribution F and any $\mathbf{t}\in\mathbb{R}^p$, we call

$$\mathbf{S}(\mathbf{X},\mathbf{t})=E\left\{\mathbf{s}(\mathbf{X}-\mathbf{t})\mathbf{s}(\mathbf{X}-\mathbf{t})^\top\right\} \tag{16.12}$$

the *spatial sign covariance matrix* with location $\mathbf{t}$ of the random vector $\mathbf{X}$. For a p-variate sample $\mathbb{X}_n=(\mathbf{X}_1,\ldots,\mathbf{X}_n)^\top$, we call

$$\hat{\mathbf{S}}_n(\mathbb{X}_n,\mathbf{t})=\frac{1}{n}\sum_{i=1}^{n}\mathbf{s}(\mathbf{X}_i-\mathbf{t})\mathbf{s}(\mathbf{X}_i-\mathbf{t})^\top \tag{16.13}$$

the (empirical) spatial sign covariance matrix with location $\mathbf{t}$ of the sample $\mathbb{X}_n$. Unless stated otherwise, we will assume in the following that the data are independent with the same distribution F. The natural location parameter $\mathbf{t}$ is the spatial median $\boldsymbol{\mu}(\mathbf{X})$, cf. (16.2), since (under very mild regularity conditions on F)

$$E\{\mathbf{s}(\mathbf{X}-\boldsymbol{\mu}(\mathbf{X}))\}=\mathbf{0},$$

and hence $\mathbf{S}(\mathbf{X}, \boldsymbol{\mu}(\mathbf{X}))$ is indeed the covariance matrix of the spatial sign of the centered random vector $\mathbf{X} - \boldsymbol{\mu}(\mathbf{X})$. Throughout the paper, we assume $\boldsymbol{\mu}(\mathbf{X})$ to be known and, without further loss of generality, assume it to be zero and only consider the theoretical and the empirical spatial sign covariance matrix with location zero, i.e.,

$$\mathbf{S}(\mathbf{X}, \mathbf{0}) = E\{\mathbf{s}(\mathbf{X})\mathbf{s}(\mathbf{X})^\top\} \qquad \text{and} \qquad \hat{\mathbf{S}}_n(\mathbb{X}_n, \mathbf{0}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(\mathbf{X}_i)\mathbf{s}(\mathbf{X}_i)^\top,$$

which we abbreviate by $\mathbf{S}(\mathbf{X})$ and $\hat{\mathbf{S}}_n$, respectively. The term *spatial sign covariance matrix* has been introduced by Visuri et al. (2000), but also earlier papers consider estimates based solely on the directions of data to analyze multivariate dependence (e.g. Locantore et al. 1999; Marden 1999).

In the following, we will frequently refer to the elliptical model, which is characterized as follows. A continuous distribution F on $\mathbb{R}^p$ is said to be *elliptical* if it has a Lebesgue-density f of the form

$$f(\mathbf{x}) = \det(\mathbf{V})^{-1/2} g\{(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})\} \tag{16.14}$$

for some $\boldsymbol{\mu} \in \mathbb{R}^p$ and symmetric, positive definite $p \times p$ matrix $\mathbf{V}$. The class of all continuous, elliptical distributions F on $\mathbb{R}^p$ having these parameters is denoted by $\mathscr{E}_p(\boldsymbol{\mu}, \mathbf{V})$. The location parameter $\boldsymbol{\mu}$ coincides with the spatial median and is henceforth assumed to be zero. We call $\mathbf{V}$ the *shape matrix*, since it describes the shape of the elliptical contour lines of the density f. The shape is unique only up to scale, that is, $\mathscr{E}_p(\boldsymbol{\mu}, \mathbf{V}) = \mathscr{E}_p(\boldsymbol{\mu}, c\mathbf{V})$ for any $c > 0$. In order to avoid this ambiguity, some authors use the term *shape matrix* exclusively for a suitably normalized representative of this equivalence class of proportional matrices, where $\det(\mathbf{V}) = 1$ is a common choice (Paindaveine 2008; Frahm 2009). For our purposes, it is more convenient to normalize the shape matrix by its trace, and we define $\mathbf{V}_0 = \mathbf{V}/\operatorname{tr}(\mathbf{V})$. Then for any $\mathbf{X} \in \mathscr{E}_p(\boldsymbol{\mu}, \mathbf{V})$ with finite second moments, we have $\boldsymbol{\Sigma}_0(\mathbf{X}) = \mathbf{V}_0(\mathbf{X})$.

The ad-hoc robustification by considering the directions $\mathbf{s}(\mathbf{X}_i)$ of the data instead of the data $\mathbf{X}_i$ themselves has some intriguing advantages and leads to a number of nice properties of the resulting scatter estimator $\hat{\mathbf{S}}_n$. Since the function $\mathbf{v}$, cf. (16.4), is bounded, asymptotic results for $\hat{\mathbf{S}}_n$ do not require any moments. The influence of any gross error is bounded, and hence $\hat{\mathbf{S}}_n$ has a bounded influence function. It has furthermore an asymptotic breakdown point of $1/2$ (Croux et al. 2010). Moreover, when sampling from the elliptical model, the distribution of $\hat{\mathbf{S}}_n$ does not depend upon the function g in (16.14). For instance, consider the two models

$$\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{V}) \qquad \text{and} \qquad \mathbf{Y} \sim t_{1,p}(\mathbf{0}, \mathbf{V})$$

for some positive definite $p \times p$ matrix $\mathbf{V}$, where $t_{1,p}$ denotes the p-variate elliptical t_1 (or Cauchy) distribution (cf. e.g. Bilodeau and Brenner 1999, p. 208). Then $\mathbf{S}(\mathbf{X}) =$

$\mathbf{S}(\mathbf{Y})$, and the respective estimators based on independent samples of size n from these two models have exactly the same distribution for every $n \in \mathbb{N}$.[1]

The main advantage, however, of $\hat{\mathbf{S}}_n$ is the computational simplicity, which must be contrasted to the tremendous computational costs of other highly robust, multivariate scatter estimators that offer a comparable degree of efficiency and robustness. Another example of a likewise simple approach is the *marginal sign covariance matrix*. It is based on the *marginal signs* of the data, i.e. the p-dimensional vectors that are composed of the univariate signs of each component. This estimator has a rather low efficiency.

But after all the praise, there is of course a drawback, and this concerns the question what $\hat{\mathbf{S}}_n$ actually estimates. Again, ignoring the location estimation, $\hat{\mathbf{S}}_n$ is apparently asymptotically normal and strongly consistent for $\mathbf{S}(\mathbf{X})$. So the question remains, what the connection between $\mathbf{S}(\mathbf{X})$ and $\boldsymbol{\Sigma}(\mathbf{X})$ is, and the answer is indeed non-trivial, as we will see below.

Much attention has been paid to affine equivariant scatter estimation. A $p \times p$ variate scatter functional $\mathbf{W}(\cdot)$ is said to be *proportionally affine equivariant* if

$$\mathbf{W}(\mathbf{AX} + \mathbf{b}) \propto \mathbf{AW}(\mathbf{X})\mathbf{A}^\top \tag{16.15}$$

for any $\mathbf{b} \in \mathbb{R}^p$ and any full rank $p \times p$ matrix $\mathbf{A}$, where $\propto$ denotes proportionality. For instance, the covariance matrix $\boldsymbol{\Sigma}$ fulfills (16.15) with equality. Affine equivariance is a very handy property. It implies that any scatter functional $\mathbf{W}$ satisfying (16.15) evaluated at an elliptical vector $\mathbf{X}$ is a multiple of the shape matrix, i.e. $\mathbf{W}(\mathbf{X}) \propto \mathbf{V}(\mathbf{X})$, cf. e.g. Tyler (1982).

The spatial sign covariance matrix lacks this property. It satisfies (16.15) only for orthogonal matrices $\mathbf{A}$, which is then referred to as *orthogonal equivariance*. The exact connection between $\mathbf{S}(\mathbf{X})$ and $\mathbf{V}(\mathbf{X})$ (up to proportionality) seems to be known only for $p = 2$ (Dürre et al. 2015). For general p, we have the following result.

Proposition 16.1 *Let* $\mathbf{X} \sim F \in \mathcal{E}_p(\mathbf{0}, \mathbf{V})$ *and let* $\mathbf{V} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ *be an eigenvalue decomposition of* $\mathbf{V}$*, where* $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ *with* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$*. Then*

$$\mathbf{S}(\mathbf{X}) = \mathbf{U}\boldsymbol{\Delta}\mathbf{U}^\top,$$

where $\boldsymbol{\Delta} = \operatorname{diag}(\delta_1, \ldots, \delta_p)$ *with* $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_p > 0$*, and* $\delta_i = \delta_j$ *if and only if* $\lambda_i = \lambda_j$ *for* $1 \leq i < j \leq p$*.*

This result is mentioned, e.g., in Magyar and Tyler (2014). It holds not only for elliptical distributions. For a proof of a more general version, see Visuri (2001). Proposition 16.1 states that the shape matrix of an elliptical distribution and the corresponding spatial sign covariance matrix share the same eigenvectors (with

[1]The exact, non-asymptotic equality in distribution does generally not hold if the location is estimated.

the usual ambiguities of eigenvectors in case of multiple eigenvalues), and that the eigenvalues are ordered accordingly. The main implication is that within the elliptical model there is a one-to-one connection between the spatial sign covariance matrix $\mathbf{S}(\mathbf{X})$ and the normalized shape matrix $\mathbf{V}_0(\mathbf{X})$ in the sense that, for any two elliptical vectors $\mathbf{X}$ and $\mathbf{Y}$, $\mathbf{S}(\mathbf{X}) = \mathbf{S}(\mathbf{Y})$ if and only if $\mathbf{V}_0(\mathbf{X}) = \mathbf{V}_0(\mathbf{Y})$. Thus, Proposition 16.1 provides the validity of our test based on

$$\hat{T}_{S,0} = \max_{1\le k\le n} \boldsymbol{\Phi}_k^\top \hat{\boldsymbol{\Omega}}_{n,0}^{-1} \boldsymbol{\Phi}_k \qquad \text{with} \qquad \boldsymbol{\Phi}_k = \tilde{\mathbf{I}}_m \operatorname{vech}\left\{ \frac{k}{\sqrt{n}} (\hat{\mathbf{S}}_k - \hat{\mathbf{S}}_n) \right\},$$

where $\hat{\mathbf{S}}_k$ denotes the sample spatial sign covariance matrix computed from $\mathbf{X}_1, \ldots, \mathbf{X}_k$, for testing the hypothesis H_0, cf. (16.3). A change in $\boldsymbol{\Sigma}_0(\mathbf{X})$ means a change in $\mathbf{S}(\mathbf{X})$ and vice versa. This holds true at least in the elliptical model, and it seems plausible that this extends to many "common data situations," although counterexamples can be constructed.

At this point it should be noted that, under ellipticity, there is generally not a one-to-one connection between $\mathbf{S}(\mathbf{X})$ and the correlation matrix $\mathbf{R}(\mathbf{X})$. For instance, the shape matrices

$$\mathbf{V}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \text{and} \qquad \mathbf{V}_2 = \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix}$$

both lead to the same correlation matrix but to spatial sign covariance matrices

$$\mathbf{S}_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix} \qquad \text{and} \qquad \mathbf{S}_2 = \begin{pmatrix} 10/11 & 0 \\ 0 & 1/11 \end{pmatrix},$$

vgl. Dürre et al. (2015).

An interesting question, which relates to the performance of the proposed change-point test, is that of the efficiency of the estimator $\hat{\mathbf{S}}_n$. Usually, the efficiencies of tests are hard to derive analytically, and comparing the efficiencies of the estimators they are based upon generally serves as a useful proxy for comparing the efficiencies of the respective tests. As we have noted above, the spatial sign covariance matrix can, except for certain special cases, not be expressed explicitly as a function of the parameters of the elliptical model. We encounter the same difficulty when analyzing its asymptotic variance. An explicit expression for the asymptotic covariance matrix of $\hat{\mathbf{S}}_n$ at elliptical distributions is only known for $p = 2$ (Dürre et al. 2015). For general p, explicit results can be given for the special case of sphericity.

Proposition 16.2 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be an independent sample from $F \in \mathscr{E}_p(\mathbf{0}, \mathbf{I}_p)$. Then $\hat{\mathbf{S}}_n$ is strongly consistent for $\mathbf{I}_p/p$ and asymptotically normal with asymptotic covariance matrix $ASV(\operatorname{vech}(\hat{\mathbf{S}}_n)) = 2/\{p(p+2)\}\mathbf{C}_p$, where*

$$\mathbf{C}_p = \mathbf{D}_p^+ (\mathbf{D}_p^+)^\top - p^{-1} \operatorname{vech}(\mathbf{I}_p) \left\{\operatorname{vech}(\mathbf{I}_p)\right\}^\top,$$

and $\mathbf{D}_p^+$ *is the Moore–Penrose inverse of the duplication matrix* $\mathbf{D}_p$ *(cf. Magnus and Neudecker 1999, p. 49). If furthermore* F *has finite fourth moments, then* $\hat{\boldsymbol{\Sigma}}_{0,n} = \hat{\boldsymbol{\Sigma}}_n / \operatorname{tr}(\hat{\boldsymbol{\Sigma}}_n)$ *is strongly consistent for* $\mathbf{I}_p/p$ *and asymptotically normal with asymptotic covariance matrix* $ASV(\operatorname{vech}(\hat{\boldsymbol{\Sigma}}_{0,n})) = 2p^{-2}(1+\kappa/3)\mathbf{C}_p$*, where* κ *denotes the excess kurtosis of any margin of* F*.*

The matrix $\mathbf{D}_p^+(\mathbf{D}_p^+)^\top$ is a diagonal matrix with each diagonal element being either 1 or 1/2. Also note that, although the symmetry redundancy has been removed in this notation, the matrix $\mathbf{C}_p$ is not of full rank, since both estimators have always trace 1. The result for the spatial sign covariance matrix can be found in Sirkiä et al. (2009). The result for $\hat{\boldsymbol{\Sigma}}_{0,n}$ can be deduced from the general form of the asymptotic covariance matrix of the sample covariance matrix at elliptical distributions, as given, e.g., in Bilodeau and Brenner (1999, p. 213), by applying the delta method to the function $\mathbf{A} \mapsto \mathbf{A}/\operatorname{tr}(\mathbf{A})$. Proposition 16.2 implies that, at the multivariate standard normal model, where $\kappa = 0$, the spatial sign covariance matrix $\hat{\mathbf{S}}_n$ is a more efficient estimator for $\mathbf{V}_0$ than the corresponding maximum-likelihood-based estimator $\hat{\boldsymbol{\Sigma}}_{0,n}$. This result may seem odd, apparently contradicting the well-known asymptotic optimality of maximum likelihood estimation.

However, $\hat{\mathbf{S}}_n$ does not consistently estimate $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}/\operatorname{tr}(\boldsymbol{\Sigma})$ within the parametric model $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In fact, for $\mathbf{X} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \not\propto \mathbf{I}_p$, the spatial sign covariance matrix $\mathbf{S}(\mathbf{X})$ is in some sense closer to $\mathbf{I}_p/p$ than $\boldsymbol{\Sigma}_0$: the eigenvalues of $\mathbf{S}(\mathbf{X})$ tend to be closer together than the eigenvalues of $\boldsymbol{\Sigma}_0(\mathbf{X})$. With this in mind, Proposition 16.2 is not more surprising than noting that $(2n)^{-1}\sum_{i=1}^n X_i$ is a more efficient location estimator than $n^{-1}\sum_{i=1}^n X_i$ at $N(0,1)$.

With $\hat{\mathbf{S}}_n$ and $\hat{\boldsymbol{\Sigma}}_n$ estimating different quantities, it is somewhat difficult to compare their asymptotic efficiencies. One possibility is to focus on those aspects of the estimators that are compatible, in this case, their eigenvectors. Magyar and Tyler (2014) study the asymptotic efficiencies of their eigenprojections. By virtue of Proposition 16.1, the eigenvectors of $\mathbf{S}(\mathbf{X})$ and $\boldsymbol{\Sigma}(\mathbf{X})$ are the same, and by virtue of the continuous mapping theorem, the corresponding sample eigenvectors based on $\hat{\mathbf{S}}_n$ and $\hat{\boldsymbol{\Sigma}}_n$, respectively, are thus consistent for the same quantity. We will not go into detail here, but refer the reader to Magyar and Tyler (2014). Their results can be summarized as follows: The asymptotic variance of the eigenprojections of $\hat{\mathbf{S}}_n$ at an elliptical distribution $F \in \mathscr{E}_p(\mathbf{0}, \mathbf{V})$ depends strongly on the ratio of the eigenvalues $\lambda_1, \ldots, \lambda_p$ of $\mathbf{V}$ and can become arbitrarily large if $\lambda_1, \ldots, \lambda_p$ differ largely in magnitude. The spherical case, i.e., $\lambda_1 = \ldots = \lambda_p$, constitutes the best case for the spatial sign covariance matrix.

So far we have seen that the transition to spatial signs provides a simple and computationally very appealing way of robustly analyzing multivariate dependencies. The price to pay, however, is an unsatisfactory efficiency under strongly shaped models (i.e., elliptical distributions with strongly squeezed or stretched elliptical contours, where the eigenvalues of the shape matrix differ greatly in size). Both can, in some sense, be attributed to the lack of affine equivariance of the spatial signs, and the question naturally arises, if this can be improved upon. In fact, the question of devising affine equivariant siblings of the spatial sign covariance matrix has been

dealt with by several authors, and we would like to mention two closely related concepts: the Tyler matrix 1987 and the Oja sign covariance matrix (Visuri et al. 2000; Ollila et al. 2003).

Tyler's scatter estimator is defined as the solution $\hat{\mathbf{V}}_n$ to

$$\hat{\mathbf{V}}_n = \frac{p}{n} \sum_{i=1}^{n} \frac{\mathbf{X}_i \mathbf{X}_i^\top}{\mathbf{X}_i^\top \hat{\mathbf{V}}_n^{-1} \mathbf{X}_i} \tag{16.16}$$

which satisfies $\mathrm{tr}(\hat{\mathbf{V}}_n) = 1$, where, for ease of exposition, we assume again the spatial median of the data to be known to be zero. Apparently, if $\hat{\mathbf{V}}_n$ is a solution to (16.16), then so is $c\hat{\mathbf{V}}_n$ for any $c > 0$. Thus, additional standardization is required to render the solution unique. We choose here, in analogy to the spatial sign covariance matrix, $\mathrm{tr}(\hat{\mathbf{V}}_n) = 1$, but $\mathrm{tr}(\hat{\mathbf{V}}_n) = p$ or $\det(\hat{\mathbf{V}}_n) = 1$ are more common. It is also easily checked that any solution to (16.16) is affine equivariant in the sense of (16.15), regardless of which specific standardization is chosen.

The similarity between Tyler's estimator and the spatial sign covariance matrix is evident: instead of dividing each observation by its Euclidean length before computing the covariance, each observation is divided by its Mahalanobis with respect to $\hat{\mathbf{V}}_n$. Equally evident is the recursive nature of this construction, and the computation of $\hat{\mathbf{V}}_n$ requires an iterative algorithm. One possibility is indeed to iterate the spatial sign covariance matrix computation, i.e., to alternate between

$$\mathbf{S} \leftarrow \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{X}_i \mathbf{X}_i^\top}{\mathbf{X}_i^\top \mathbf{S}^{-1} \mathbf{X}_i} \qquad \text{and} \qquad \mathbf{S} \leftarrow \mathbf{S}/\,\mathrm{tr}(\mathbf{S})$$

until convergence is reached, starting with $\mathbf{S}_0 = \mathbf{I}_p$. The estimate obtained by stopping this algorithm after a fixed number of steps has been called the *k-step spatial sign covariance matrix* by Croux et al. (2010). The data divided by their Mahalanobis distances $\mathbf{d}_i = \{\mathbf{X}_i^\top \hat{\mathbf{V}}_n \mathbf{X}_i\}^{1/2}$ with respect to the Tyler matrix $\hat{\mathbf{V}}_n$ can be interpreted as affine equivariant spatial signs.

Another multivariate sign which obeys some form of affine equivariance is the *Oja sign*. Let

$$Q_{p-1} = \big\{ q = \{i_1, \ldots, i_{p-1}\} \;\big|\; 1 \le i_1 < \ldots < i_{p-1} \le n \big\}, \quad 1 \le p \le n,$$

be the set of all subsets of $\{1, \ldots, n\}$ with $p-1$ elements, and $N_{p-1} = |\,Q_{p-1}| = \binom{n}{p-1}$. Then, again assuming the data to be already suitably centered, the Oja sign of the point $\mathbf{x} \in \mathbb{R}^p$ with respect to the sample $\mathbb{X}_n = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top$ is defined as

$$\mathbf{o}_{\mathbb{X}_n}(\mathbf{x}) = \frac{1}{N_{p-1}} \sum_{q \in Q_{p-1}} \nabla_{\mathbf{x}} \Big| \det(\mathbf{X}_{i_1} \;\ldots\; \mathbf{X}_{i_{p-1}} \;\mathbf{x}) \Big|, \tag{16.17}$$

where $\nabla_{\mathbf{x}}$ denotes the gradient with respect to $\mathbf{x}$. The *Oja sign covariance matrix with location* $\mathbf{0}$ of the sample $\mathbb{X}_n$ is then defined as

$$\hat{\mathbf{O}}_n = \hat{\mathbf{O}}_n(\mathbb{X}_n, \mathbf{0}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{o}_{\mathbb{X}_n}(\mathbf{X}_i)\mathbf{o}_{\mathbb{X}_n}(\mathbf{X}_i)^\top. \tag{16.18}$$

The Oja sign allows the following geometric interpretation: In dimension two, $\mathbf{o}_{\mathbb{X}_n}(\mathbf{x})$ is the average of $N_1 = n$ vectors $\mathbf{v}_i$, $1 \le i \le n$, where $\mathbf{v}_i$ is perpendicular to $\mathbf{X}_i$, has the same length as $\mathbf{X}_i$, and points towards $\mathbf{x}$. In dimension three, $\mathbf{o}_{\mathbb{X}_n}(\mathbf{x})$ is the average of $N_2 = \binom{n}{2}$ vectors $\mathbf{v}_q$, where each is perpendicular to the plane spanned by $\mathbf{X}_i$ and $\mathbf{X}_j$, $q = \{i, j\}$, points towards $\mathbf{x}$, and its length is twice the area of the triangle formed by $\mathbf{X}_i$, $\mathbf{X}_j$ and $\mathbf{0}$.

Also note that the Oja sign is a proper generalization of the univariate sign: recalling that $\{1, \dots, n\}$ has one subset of size zero (the empty set), the sum in (16.17) has one summand for $p = 1$, and we get $\mathbf{o}_{\mathbb{X}_n}(x) = d|\det(x)|/dx$, which is another way of writing down the univariate sign function $s(x) = \mathbb{1}_{\{x>0\}} - \mathbb{1}_{\{x<0\}}$.

Further, it can be seen from (16.17) that

$$\mathbf{o}_{\mathbb{X}_n\mathbf{A}^\top}(\mathbf{A}\mathbf{x}) = \det(\mathbf{A})\mathbf{A}^{-1}\mathbf{o}_{\mathbb{X}_n}(\mathbf{x})$$

for any full rank matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$, where $\mathbb{X}\mathbf{A}^\top = (\mathbf{A}\mathbf{X}_1, \dots, \mathbf{A}\mathbf{X}_n)^\top$ denotes the data sample obtained from $\mathbf{X}$ by applying the linear transformation $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$. This inverse affine equivariance of the Oja sign implies that the *inverse* of the population version of the Oja sign covariance matrix (see Ollila et al. 2003) is affine equivariant in the sense of (16.15).

So, the Tyler matrix $\hat{\mathbf{V}}_n$ as well as the trace-standardized inverse of the Oja sign covariance matrix, i.e. $\hat{\mathbf{O}}_n^{-1}/\operatorname{tr}(\hat{\mathbf{O}}_n^{-1})$, are both affine equivariant and hence, under appropriate regularity conditions, consistent for $\mathbf{V}_0$ under ellipticity. Furthermore, their asymptotic covariance matrices are proportional to that of $\hat{\boldsymbol{\Sigma}}_{0,n} = \hat{\boldsymbol{\Sigma}}_n/\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_n)$, which makes it much easier to compare them efficiency-wise. It turns out that both estimators have a rather good efficiency also at the normal model as compared to $\hat{\boldsymbol{\Sigma}}_{0,n}$ (cf. Tyler 1987; Ollila et al. 2003). This efficiency is independent of the shape of the elliptical distribution, implying in particular that there is no loss under strong shapedness as it is the case for the spatial sign covariance matrix.

But there are a number of drawbacks. First of all, $\hat{\mathbf{V}}_n$ and $\hat{\mathbf{O}}_n$ both require $n \ge p + 1$ data points in order to be defined, and that is in some sense true for all robust, affine equivariant scatter estimators (Tyler 2010). Second, they have weaker robustness properties: The Oja sign covariance matrix has an unbounded influence function and requires finite second moments to be asymptotically normal. The Tyler matrix has a bounded influence function and provides, for practical purposes, a similarly high degree of outlier resistance as the spatial sign covariance matrix, but lacks, for instance, the high breakdown point property (Dümbgen and Tyler 2005). Third, and most severely, they have a much higher computational cost than the spatial sign covariance matrix, which makes them unfeasible for

applications in high-dimensional change-point detection. While there is an explicit expression (16.17, 16.18) for the Oja sign covariance matrix, making it, contrary to the Tyler matrix, computable in a fixed number of steps, the algorithmic comparison is clearly in favor of the Tyler matrix: the iterative algorithm described above converges quickly, and Tyler's estimator can be computed basically in any dimension where matrix inversion is still numerically practicable. The computation of the Oja sign covariance matrix, on the other hand, necessitates the evaluation of $\binom{n}{p-1}$ hyperplanes in $\mathbb{R}^p$, which is clearly prohibitive for many data situations where change detection is of interest. The R-package OjaNP (Fischer et al. 2014) therefore provides also a non-exact, subsampling-based version of the Oja sign covariance matrix.

16.4 Simulation Results

We compare the change-point test introduced in Sect. 16.2, based on the spatial sign covariance matrix, to the sample-covariance-based test by Aue et al. (2009), cf. (16.8), in terms of size and power. Both test statistics are equally fast to compute. Throughout, we use 1000 repetitions, dimensions $p = 6$ and $p = 30$ and various sample sizes ranging from $n = 200$ to $n = 2000$. We only consider one-change alternatives, where a change in the dependence structure occurs in the middle of the observed period, and the data are stationary before and after the change-point. We consider three different scenarios concerning the distribution of the data.

(a) The p-variate observations $\mathbf{X}_i$ are independent, each being elliptically distributed. The elliptical distributions have equal marginal variances.
(b) The observations are serially dependent, according to a multivariate GARCH model with normal innovations. Again, the marginal variances are equal.
(c) The observations are independent, as in (a), but different transformations are applied to the margins.

The three scenarios are explained in detail below.

(a) Independent, Elliptically Distributed Observations The observations $\mathbf{X}_1, \ldots, \mathbf{X}_{\lfloor n/2 \rfloor}$ are identically distributed with mean zero and covariance matrix $\boldsymbol{\Sigma}_1$, and $\mathbf{X}_{\lfloor n/2 \rfloor + 1}, \ldots, \mathbf{X}_n$ are identically distributed also with mean zero but covariance matrix $\boldsymbol{\Sigma}_2$. We use the multivariate normal distribution and the elliptical $t_{5,p}$ distribution with covariance matrices $\boldsymbol{\Sigma}_1 = \mathbf{I}_p$ and $\boldsymbol{\Sigma}_2$ being defined element-wise as

$$(\boldsymbol{\Sigma}_2)_{i,j} = \begin{cases} 1 & \text{for } i = j, \\ h & \text{for } i \neq j, \end{cases} \tag{16.19}$$

for some $-1/(p-1) < h < 1$ and all $1 \leq i, j \leq p$. Any choice of h within that range leads to a positive definite matrix, as can be deduced from results about circulant

Table 16.1 Comparison of change-point tests at independent, elliptically distributed observations with equal marginal variances

Distribution			Normal					t_5				
p	n	Estimator	Size of change h									
			0.0	0.05	0.1	0.2	0.4	0.0	0.05	0.1	0.2	0.4
6	200	$\hat{\boldsymbol{\Sigma}}_n$	2	3	6	30	97	< 1	1	2	9	69
		$\hat{\mathbf{S}}_n$	3	4	7	31	93	3	3	9	30	94
		$\hat{\boldsymbol{\Sigma}}_n$ [a]	5	7	13	46	99	5	8	9	29	91
		$\hat{\mathbf{S}}_n$ [a]	5	6	12	39	96	5	6	12	39	97
	500	$\hat{\boldsymbol{\Sigma}}_n$	4	8	33	97	100	< 1	3	14	65	100
		$\hat{\mathbf{S}}_n$	4	8	29	90	100	5	8	30	91	100
	1000	$\hat{\boldsymbol{\Sigma}}_n$	4	23	80	100	100	2	9	40	98	100
		$\hat{\mathbf{S}}_n$	4	20	66	100	100	5	17	66	100	100
30	500	$\hat{\boldsymbol{\Sigma}}_n$	0	0	0	0	0	0	0	0	0	0
		$\hat{\mathbf{S}}_n$	0	0	0	0	0	0	0	0	0	0
		$\hat{\boldsymbol{\Sigma}}_n$ [a]	5	8	11	30	93	5	9	12	16	47
		$\hat{\mathbf{S}}_n$ [a]	5	11	18	55	93	5	13	22	46	95
	1000	$\hat{\boldsymbol{\Sigma}}_n$	< 1	13	63	100	100	< 1	2	16	66	99
		$\hat{\mathbf{S}}_n$	< 1	20	83	100	100	< 1	21	87	100	100
		$\hat{\boldsymbol{\Sigma}}_n$ [a]	5	54	96	100	100	5	43	82	98	100
		$\hat{\mathbf{S}}_n$ [a]	5	62	99	100	100	5	62	99	100	100

Rejection frequencies (%) at the 5 % significance level, based on 1000 repetitions
[a]With empirical critical values

matrices (cf. e.g. Gray 2006). We let h range from 0 to 0.4, cf. Table 16.1. The choice $h = 0$ corresponds to the null hypothesis of no change. Note that, also under the alternatives, the p univariate marginal processes are stationary. The $t_{\nu,p}$ distribution has finite α-moments for any $\alpha < \nu$. In particular, for $\nu > 4$, the margins of a $t_{\nu,p}$ distribution have excess kurtosis $\kappa = 6/(\nu - 4)$. In the independence scenarios (a) and (c), the test statistic $\hat{T}_{S,0}$ of Theorem 16.1 is used with the long-run variance estimator $\hat{\boldsymbol{\Omega}}_{n,0}$, cf. (16.6), which does not account for potential serial dependence. Analogously, we take as long-run variance estimator for vech $\hat{\boldsymbol{\Sigma}}_n$

$$\hat{\boldsymbol{\Gamma}}_{n,0} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{vech}(\mathbf{X}_i\mathbf{X}_i^\top).$$

In Table 16.1, the rejection frequencies at the 5 % significance level based on the asymptotic critical values given by Theorem 16.1 are displayed for $p = 6$ and $p = 30$ and several values of n and h. As far as size is concerned, we find that, in all situations, both tests keep the size under the null hypothesis but tend to be conservative, with a clear superiority of the spatial-sign-based test. In cases where both tests are largely below the nominal 5 % level, sample-size adjusted rejection frequencies based on the empirically determined critical values are also

given (indicated by superscript "a"). In particular, the zeroes for the combination $(p, n) = (30, 500)$ are no typo. In this situation, the sample size is just little above the parameter dimension of $m = 464$. This illustrates that in such a situation, the distributions of the test statistics are yet far away from their limit. We further observe that the spatial-sign-based test is better than the variance-based test at the $t_{5,p}$ distribution throughout, but also for $n = 30$ under normality. The invariance of the spatial sign covariance matrix with respect to the elliptical generator g, cf. (16.14), appears to extend to the corresponding change-point test.

(b) Multivariate GARCH We employ the constant conditional correlation (CCC) GARCH(1,1) model, introduced by Bollerslev (1990). Let

$$\mathbf{X}_i = \boldsymbol{\tau}_i \circ \mathbf{Y}_i, \quad \text{with} \quad \begin{pmatrix} \tau_{i,1}^2 \\ \vdots \\ \tau_{i,p}^2 \end{pmatrix} = \begin{pmatrix} 0.1 \\ \vdots \\ 0.1 \end{pmatrix} + 0.1 \begin{pmatrix} X_{i,1}^2 \\ \vdots \\ X_{i,p}^2 \end{pmatrix} + 0.84 \begin{pmatrix} \tau_{i,1}^2 \\ \vdots \\ \tau_{i,p}^2 \end{pmatrix}, \quad i \in \mathbb{Z},$$

where $\boldsymbol{\tau} = (\tau_{i,1}, \ldots, \tau_{i,p})^\top$, $\mathbf{Y}_i = (Y_{i,1}, \ldots, Y_{i,p})^\top$, and $\circ$ denotes the Hadamard product, i.e., component-wise multiplication. The innovations $\mathbf{Y}_i$ are independent and normally distributed with mean zero and covariance matrix $\boldsymbol{\Sigma}_1 = \mathbf{I}_p$ before and covariance matrix $\boldsymbol{\Sigma}_2$, cf. (16.19), after the change in the middle of the observed period. Thus, as in scenario (a), the univariate marginal processes are stationary (given a sufficient burn-in period) under the null as well as under the alternatives. Their stationary distributions possess heavier than normal tails. However, contrary to (a), the data $\mathbf{X}_i$ are neither elliptical nor do they possess the same covariance as $\mathbf{Y}_i$.

We estimate the long-run variance $\boldsymbol{\Omega}$ by $\hat{\boldsymbol{\Omega}}_{n,b_n}$, cf. (16.9). Similarly, we estimate the long-run variance of the sample-variance-based test by

$$\hat{\boldsymbol{\Gamma}}_{n,b_n} = \frac{1}{n} \sum_{k=\lfloor -b_n \rfloor}^{\lceil b_n \rceil} \left(1 - \frac{|k|}{b_n + 1}\right) \sum_{j=1}^{n-|k|} \operatorname{vech}(\mathbf{X}_j \mathbf{X}_j^\top) \{\operatorname{vech}(\mathbf{X}_{j+k} \mathbf{X}_{j+k}^\top)\}^\top.$$

In both cases, we take $b_n = 1.5n^{1/3}$. The choice of b_n considerably affects the behavior of the test. Choosing it too low may cause the test to exceed the nominal size level, choosing it too large will result in a loss of power. We have tried several values for b_n, and our choice proved to be a good compromise. How to select the bandwidth in an optimal, data-adaptive way is an important and delicate question, which goes beyond the scope of this paper. Given the similar structure of the tests and long-run variance estimates, choosing b_n equal for both tests allows a fair comparison. The results for scenario (b) are summarized in Table 16.2. The convergence of the test statistic to the asymptotic distribution under the null is significantly slower than in the independent case. For $p = 30$, we are far away from the asymptotic distribution even for $n = 2000$. So we also include the rejection frequencies for the empirically determined critical values. They are not

Table 16.2 Comparison of change-point tests at multivariate GARCH processes with normal innovations

Critical values			Asymptotic					Empirical				
p	n	Estimator	Size of change h									
			0.0	0.05	0.1	0.2	0.4	0.0	0.05	0.1	0.2	0.4
6	500	$\hat{\boldsymbol{\Sigma}}_n$	0	0	0	0	0	5	8	13	55	100
		$\hat{\mathbf{S}}_n$	0	0	0	0	6	5	10	16	57	100
	1000	$\hat{\boldsymbol{\Sigma}}_n$	0	2	9	81	100	5	12	43	99	100
		$\hat{\mathbf{S}}_n$	1	2	10	82	100	5	13	40	99	100
	2000	$\hat{\boldsymbol{\Sigma}}_n$	3	18	90	100	100	5	28	93	100	100
		$\hat{\mathbf{S}}_n$	3	17	82	100	100	5	23	88	100	100
30	500	$\hat{\boldsymbol{\Sigma}}_n$	0	0	0	0	0	5	4	6	5	8
		$\hat{\mathbf{S}}_n$	0	0	0	0	0	5	5	8	7	11
	1000	$\hat{\boldsymbol{\Sigma}}_n$	0	0	0	0	0	5	14	26	54	98
		$\hat{\mathbf{S}}_n$	0	0	0	0	0	5	17	36	66	96
	2000	$\hat{\boldsymbol{\Sigma}}_n$	0	0	0	0	0	5	43	83	100	100
		$\hat{\mathbf{S}}_n$	0	0	0	0	0	5	52	94	100	100

Rejection frequencies (%) at the 5 % significance level, based on 1000 repetitions

marked by an asterisk as in Table 16.1, but can be found on the right-hand side of Table 16.2. It must be noted that the discrepancy between the asymptotic and the actual distribution is not an artifact of the bandwidth b_n being chosen too large. This effect also occurs if the long-run variance is assumed to be known, which is hard to derive analytically, but can easily be gotten hold of by means of simulation. Also note that Aue et al. (2009) report simulation results for $p = 4$ only.

Comparing both tests, we find that they behave rather similar, with slight advantages for the variance-based test for $p = 6$, and for the spatial-sign-based test for $p = 30$.

(c) Independent Observations with Unequal Marginal Distributions The previous setups with the equally dispersed margins are in some sense favorable for the spatial sign covariance matrix, since different univariate marginal variances will increase the shapedness (i.e., the ratio of the eigenvalues of the shape matrix $\mathbf{V}$). Therefore we consider in the following scenarios where not all variables have the same marginal variance. For demonstration purposes, we only consider $p = 6$ and $n = 1000$. Starting from the normal model considered in (A), with covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ before and after the change, we apply two different types of marginal transformations to the data. In the first setting, we multiply the first r components of the p-variate observations by 10, where r ranges from 0 to 5. This is equally done before and after the change, such that the univariate marginal distributions are again stationary also under the alternatives. The resulting joint distributions (which are, as above, different before and after the change) are strongly shaped normal distributions. The rejection frequencies of both tests are shown on the left-hand side of Table 16.3. We call this the strongly shaped scenario.

Table 16.3 Comparison of change-point tests at distributions with unequal margins

Type		Strongly shaped					Heavy-tailed				
r	Estimator	Size of change h									
		0.0	0.05	0.1	0.2	0.4	0.0	0.05	0.1	0.2	0.4
0	$\hat{\boldsymbol{\Sigma}}_n$	4	23	80	100	100	4	23	80	100	100
	$\hat{\mathbf{S}}_n$	4	20	66	100	100	4	20	66	100	100
1	$\hat{\boldsymbol{\Sigma}}_n$	4	22	82	100	100	3	15	66	100	100
	$\hat{\mathbf{S}}_n$	3	8	27	88	100	4	16	62	100	100
2	$\hat{\boldsymbol{\Sigma}}_n$	4	21	80	100	100	2	8	50	100	100
	$\hat{\mathbf{S}}_n$	4	7	30	90	100	4	16	60	100	100
3	$\hat{\boldsymbol{\Sigma}}_n$	3	22	81	100	100	2	6	36	99	100
	$\hat{\mathbf{S}}_n$	4	11	44	99	100	4	14	57	100	100
4	$\hat{\boldsymbol{\Sigma}}_n$	3	20	81	100	100	2	5	21	92	100
	$\hat{\mathbf{S}}_n$	4	13	51	100	100	4	16	54	100	100
5	$\hat{\boldsymbol{\Sigma}}_n$	3	19	82	100	100	2	2	12	69	100
	$\hat{\mathbf{S}}_n$	4	13	62	100	100	4	12	53	99	100

Rejection frequencies (%) at the 5 % significance level, based on 1000 repetitions; dimension $p = 6$; observations $n = 1000$; r number of margins transformed

In the second setting, we multiply *some* of the entries of the data vectors by 10. In the first r components of each data vector, each entry is multiplied by 10 with probability $\epsilon = 0.1$ independently of the others, and this again likewise before and after the change. Note that the resulting univariate marginal distributions are not normal, and the joint distribution in this setting is not elliptical anymore. Also note that this marginal transformation also changes the correlations and that hence the transformed sequence after the change does generally not have correlation matrix $\boldsymbol{\Sigma}_2$. The results for this set-up, which we label the heavy-tailed scenario, are summarized on the right-hand side of Table 16.3.

Although both transformations are very similar in their construction, they have quite different, opposing effects on the two tests. In the strongly shaped scenario, we see no impact on the variance-based test, but a considerable loss for the spatial-sign-based test. The power of the spatial-sign-based test is worst for $r = 1$ and increases as r increases. This observation is in line with the results of Magyar and Tyler (2014), who note that a few large eigenvalues are worse for the spatial sign covariance matrix than a few small eigenvalues. In the heavy-tailed scenario, we find that the variance-based test strongly deteriorates as r increases, while the loss of the spatial-sign-based test is very moderate. This behavior is mainly attributed to the inefficiency of the sample covariance matrix in case of a large kurtosis. Here, the marginal distribution of the transformed components is a normal scale mixture of the type $\epsilon N(0, \lambda^2) + (1 - \epsilon)N(0, 1)$, which has, for $\lambda = 10$ and $\epsilon = 0.1$, an excess kurtosis of about $\kappa = 22.3$. This family, sometimes also referred to as the *contaminated normal model*, has been used by Tukey (1960) to demonstrate the non-robustness of the sample standard deviation as a scale measure.

Finally, shortcomings of either of the tests due to unequally distributed margins (in whatever way) can be circumvented relatively easily by applying appropriate marginal transformations. This has also been noted by Aue et al. (2009) as a technique to apply their test to data without fourth moments. Changing the scale is comparably unambiguous: one divides each margin by a (preferably robust) univariate scale estimate. The effect of such a standardization on the asymptotic distribution of the spatial sign covariance matrix has been discussed by Dürre et al. (2015). However, when changing the shape of the distribution (now speaking of the *shape* of univariate densities, not elliptical contours) in order to reduce the tailedness, one has a certain liberty in choosing the specific transformation. Different transformations will affect the multivariate structure of the joint distributions in different ways and hence may alter the results of the subsequent statistical analysis, opening the door to deliberate data manipulations. It is an open research question to devise a method that chooses an appropriate data transformation in an objective, purely data-dependent way (cf. also Aue et al. 2009, Sect. 3.2). It is generally advisable to standardize the margins prior to applying the spatial-sign-based test.

16.5 Discussion and Outlook

In this paper, we presented a robust change-point test for multivariate dependence. We propose a simple adjustment to the sample-covariance based test by Aue et al. (2009), that is, to use the spatial signs of the data instead of the data itself. Thus we test for the constancy of the spatial sign covariance matrix instead of the covariance matrix, and the test is based on the sample spatial sign covariance matrix. This simple construction allows to apply well-established asymptotic tools in the proofs and yet leads to a fast-to-compute test with good robustness and efficiency properties.

A certain reluctance against the use of robust estimators in general stems from the strong focus on least-squares characteristics as descriptive parameters of distributions. For instance, the mean is used to describe central location, and any alternative location measure does, by definition, coincide with the mean at most in some submodel. This objection is of much lesser legitimacy in two-sample or change-point tests. We have established in Sect. 16.3 that there is a one-to-one connection between the spatial sign covariance matrix $\mathbf{S}(\mathbf{X})$ and the standardized covariance matrix $\boldsymbol{\Sigma}_0(\mathbf{X})$ within the elliptical model, meaning a change in one multivariate parameter implies a change in the other and vice versa. The change may be of different magnitude, but both multivariate parameters, $\boldsymbol{\Sigma}_0(\mathbf{X})$ and $\mathbf{S}(\mathbf{X})$, give each rise to change-point tests for multivariate dependence regardless of any of them being considered an important descriptive statistic of multivariate distributions.

We have also noted that, under ellipticity, there is generally no one-to-one connection between the spatial sign covariance matrix and the correlation matrix $\mathbf{R}(\mathbf{X})$. However, spatial signs can also be used to estimate $\mathbf{R}(\mathbf{X})$ under ellipticity

and hence to devise tests that specifically concern $\mathbf{R}(\mathbf{X})$. For details see Dürre et al. (2015).

The main restriction of Theorems 16.1 and 16.2 is that they assume the spatial median $\boldsymbol{\mu}(\mathbf{X})$ of the population distribution to be known. This is unrealistic but means no severe limitation of the practical applicability of the test. One may center the data beforehand by any fixed location $\mathbf{t} \in \mathbb{R}^p$ and subsequently use the results of this article to test for the constancy of the spatial sign covariance matrix with location $\mathbf{t}$, cf. (16.12). A change in the multivariate dependence structure of $\mathbf{X}$ will also show up in the incorrectly centered spatial sign covariance matrix $\mathbf{S}(\mathbf{X}, \mathbf{t})$. The change will generally be less pronounced and the test consequently less efficient, but presumably very little so if $\mathbf{t}$ is close to the actual central location $\boldsymbol{\mu}$.

However, from a theoretical point of view, it is interesting to know if and how the results given here can be extended to the unknown-location case. We conjecture that Theorem 16.2 holds also true under slightly stronger conditions if the observations $\mathbf{X}_i$ are replaced by $\mathbf{X}_i - \hat{\boldsymbol{\mu}}_n$, where $\hat{\boldsymbol{\mu}}_n$ is a location estimator based on the whole sample $\mathbb{X}_n$. The canonical estimator for that purpose is the (empirical) spatial median

$$\hat{\boldsymbol{\mu}}_n(\mathbb{X}_n) = \underset{\boldsymbol{\mu} \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \|\mathbf{X}_i - \boldsymbol{\mu}\|.$$

It is also highly robust (e.g., Lopuhaä and Rousseeuw 1991) and very fast to compute (e.g., Vardi and Zhang 2001). Its efficiency at elliptical distributions is studied in detail by Magyar and Tyler (2011). For independent observations, Dürre et al. (2014) give conditions for

$$\sqrt{n}\,\mathrm{vec}\{\hat{\mathbf{S}}_n(\mathbb{X}_n, \hat{\boldsymbol{\mu}}_n) - \mathbf{S}(\mathbf{X}, \boldsymbol{\mu})\}$$

to be asymptotically normal. This result needs to be generalized in two respects: to dependent data and to process convergence, i.e., we require

$$\left(\sqrt{n}\,t\,\mathrm{vec}\{\hat{\mathbf{S}}_{\lfloor nt \rfloor}(\mathbf{X}_{\lfloor nt \rfloor}, \hat{\boldsymbol{\mu}}_n) - \mathbf{S}(\mathbf{X}, \boldsymbol{\mu})\}\right)_{0 \le t \le 1}$$

to converge to a multivariate Brownian motion, where $\mathbb{X}_k = (\mathbf{X}_1, \ldots, \mathbf{X}_k)^\top$, $k = 1, \ldots, n$. Furthermore, consistency of the long-run variance estimator needs to be shown. Here is yet some substantial work to do and this is a topic for future research.

Acknowledgements This work was supported in part by the Collaborative Research Grant 823 of the German Research Foundation.

References

Aue, A., Hörmann, S., Horváth, L., Reimherr, M.: Break detection in the covariance structure of multivariate time series models. Ann. Stat. **37**(6B), 4046–4087 (2009)

Bilodeau, M., Brenner, D.: Theory of Multivariate Statistics. Springer Texts in Statistics. Springer, New York (1999)

Bollerslev, T.: Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model. Rev. Econ. Stat. **72**, 498–505 (1990)

Bücher, A., Kojadinovic, I., Rohmer, T., Segers, J.: Detecting changes in cross-sectional dependence in multivariate time series. J. Multivar. Anal. **132**, 111–128 (2014)

Croux, C., Dehon, C., Yadine, A.: The *k*-step spatial sign covariance matrix. Adv. Data Anal. Classif. **4**(2–3), 137–150 (2010)

de Jong, R.M., Davidson, J.: Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. Econometrica **68**(2), 407–424 (2000)

Dehling, H., Vogel, D., Wendler, M., Wied, D.: Testing for changes in the rank correlation of time series. version 1 (2012) [arXiv 1203.4871]

Dümbgen, L., Tyler, D.E.: On the breakdown properties of some multivariate M-functionals. Scand. J. Stat. **32**(2), 247–264 (2005)

Dürre, A., Vogel, D., Tyler, D.E.: The spatial sign covariance matrix with unknown location. J. Multivar. Anal. **130**, 107–117 (2014)

Dürre, D., Vogel, D., Fried, R.: Spatial sign correlation. J. Multivar. Anal. **135**, 89–105 (2015)

Fischer, D., Möttönen, J., Nordhausen, K., Vogel, D.: OjaNP: Multivariate Methods Based on the Oja Median and Related Concepts, 2014. http://CRAN.R-project.org/package=OjaNP. R package version 0.9-8

Frahm, G.: Asymptotic distributions of robust shape matrices and scales. J. Multivar. Anal. **100**(7), 1329–1337 (2009)

Galeano, P., Peña, D.: Covariance changes detection in multivariate time series. J. Stat. Plann. Inference **137**(1), 194–211 (2007)

Galeano, P., Wied, D.: Multiple break detection in the correlation structure of random variables. Comput. Stat. Data Anal. **76**, 262–282 (2014)

Gray, R.M.: Toeplitz and circulant matrices: a review. Found. Trends Commun. Inf. Theory **2**(3), 155–239 (2006)

Ibragimov, I.A.: Some limit theorems for stationary processes. Theory Probab. Appl. **7**, 349–382 (1962)

Kojadinovic, I., Quessy, J.F., Rohmer, T.: Testing the constancy of Spearman's rho in multivariate time series. Ann. Inst. Math. Stat. (2015). doi: 10.1007/s10463-015-0520-2

Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., Cohen, K.: Robust principal component analysis for functional data. Test **8**(1), 1–73 (1999)

Lopuhaä, H.P., Rousseeuw, P.J.: Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. Ann. Stat. **19**(1), 229–248 (1991)

Magnus, J.R., Neudecker, H.: Matrix Differential Calculus with Applications in Statistics and Econometrics, 2nd edn. Wiley Series in Probability and Statistics. Wiley, Chichester (1999)

Magyar, A.F., Tyler, D.E.: The asymptotic efficiency of the spatial median for elliptically symmetric distributions. Sankhya B **73**(2), 165–192 (2011)

Magyar, A.F., Tyler, D.E.: The asymptotic inadmissibility of the spatial sign covariance matrix for elliptically symmetric distributions. Biometrika **101**(3), 673–688 (2014)

Marden, J.I.: Some robust estimates of principal components. Stat. Probab. Lett. **43**(4), 349–359 (1999)

Ollila, E., Oja, H., Croux, C.: The affine equivariant sign covariance matrix: asymptotic behavior and efficiencies. J. Multivar. Anal. **87**(2), 328–355 (2003)

Paindaveine, D.: A canonical definition of shape. Stat. Probab. Lett. **78**(14), 2240–2247 (2008)

Quessy, J.-F., Saïd, M., Favre, A.-C.: Multivariate Kendall's tau for change-point detection in copulas. Can. J. Stat. **41**, 65–82 (2013)

Sirkiä, S., Taskinen, S., Oja, H., Tyler, D.E.: Tests and estimates of shape based on spatial signs and ranks. J. Nonparametric Stat. **21**(2), 155–176 (2009)

Tukey, J.W.: A survey of sampling from contaminated distributions. In: Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G., Mann, H.B. (eds.) Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling, pp. 448–485. Stanford University Press, Stanford (1960)

Tyler, D.E.: Radial estimates and the test for sphericity. Biometrika **69**, 429–436 (1982)

Tyler, D.E.: A distribution-free M-estimator of multivariate scatter. Ann. Stat. **15**, 234–251 (1987)

Tyler, D.E.: A note on multivariate location and scatter statistics for sparse data sets. Stat. Probab. Lett. **80**(17-18), 1409–1413 (2010)

Vardi, Y., Zhang, C.H.: A modified Weiszfeld algorithm for the Fermat-Weber location problem. Math. Program. Ser. A **90**, 559–566 (2001)

Visuri, S.: Array and multichannel signal processing using nonparametric statistics. Ph.D. Thesis, Helsinki University of Technology, Helsinki, Finland (2001)

Visuri, S., Koivunen, V., Oja, H.: Sign and rank covariance matrices. J. Stat. Plann. Inference **91**(2), 557–575 (2000)

Wied, D., Krämer, W., Dehling, H.: Testing for a change in correlation at an unknown point in time using an extended delta method. Econometric Theory **28**(3), 570–589 (2012)

Wied, D., Dehling, H., van Kampen, M., Vogel, D.: A fluctuation test for constant Spearman's rho with nuisance-free limit distribution. Comput. Stat. Data Anal. **76** 723–736, (2014)

Wooldridge, J.M., White, H.: Some invariance principles and central limit theorems for dependent heterogeneous processes. Econometric Theory **4**, 210–230 (1988)

Chapter 17
Tyler's M-Estimator in High-Dimensional Financial-Data Analysis

Gabriel Frahm and Uwe Jaekel

Abstract Standard methods of random matrix theory have been often applied to high-dimensional financial data. We discuss the fundamental results and potential shortcomings of random matrix theory by taking the stylized facts of empirical finance into consideration. In particular, the Marčenko–Pastur law generally fails when analyzing the spectral distribution of the sample covariance matrix if the data are generalized spherically distributed and heavy tailed. We propose Tyler's M-estimator as an alternative. Substituting the sample covariance matrix by Tyler's M-estimator resolves the typical difficulties that occur in financial-data analysis. In particular, the Marčenko–Pastur law remains valid. This holds even if the data are not generalized spherically distributed but independent and identically distributed.

Keywords Angular central Gaussian distribution • Financial data • High-dimensional data • Marčenko–Pastur law • Principal-component analysis • Random matrix theory • Spectral distribution • Tyler's M-estimator • Wigner's semicircle law

17.1 Motivation

The distribution of short-term asset returns usually exhibits heavy tails or at least leptokurtosis, tail dependence, skewness, volatility clusters or even long memory, etc. Moreover, high-frequency data generally are non-stationary, have jumps, and are strongly dependent. These stylized facts can be observed in particular for stocks, stock indices, and foreign exchange rates. Indeed, the literature on this topic is overwhelming (Bouchaud et al. 1997; Breymann et al. 2003; Ding et al. 1993;

G. Frahm (✉)
Department of Mathematics/Statistics, Helmut Schmidt University/University of the Federal Armed Forces Germany, Holstenhofweg 85, 22043 Hamburg, Germany
e-mail: frahm@hsu-hh.de

U. Jaekel
Department of Mathematics and Technology, University of Applied Sciences Koblenz, RheinAhrCampus, Südallee 2, 53424 Remagen, Germany
e-mail: jaekel@hs-koblenz.de

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_17

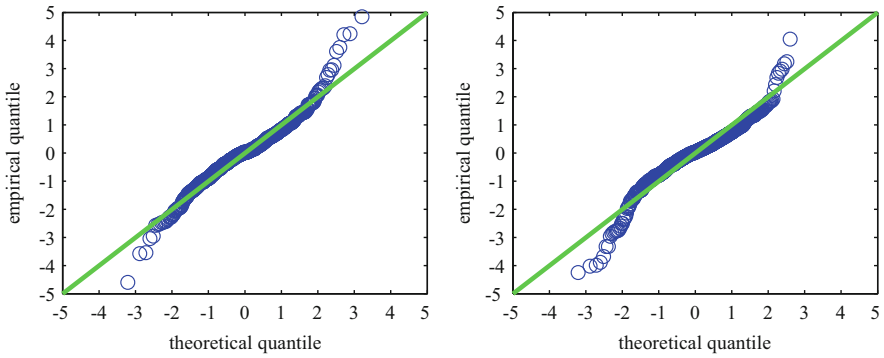

Fig. 17.1 Normal Q-Q plots of daily log-returns on OMX Helsinki 25 (*left*) and DAX 30 (*right*) from 2007-01-03 to 2009-12-31 ($n = 756$)

Eberlein and Keller 1995; Embrechts et al. 1997; Engle 1982; Fama 1965; Junker and May 2005; Mandelbrot 1963; McNeil et al. 2005; Mikosch 2003, etc.).

Figure 17.1 shows normal Q-Q plots of daily log-returns on the OMX Helsinki 25 and DAX 30 from 2007-01-03 to 2009-12-31.[1] Hence, the chosen period covers the financial crisis 2007–2009. During this period we can observe $n = 756$ log-returns and the given Q-Q plots clearly indicate that the normal-distribution hypothesis is inappropriate. More precisely, the probability of extremes is much higher than suggested by the normal distribution.

Figure 17.2 shows the joint distribution of the log-returns considered above. We can observe the following effects in the scatter plot:

- The central region of the distribution seems to be elliptically contoured.
- In the margins, the joint distribution of asset returns is asymmetric.
- There are a number of outliers or extreme values.
- Extremes typically occur simultaneously.

The fact that extreme asset returns typically occur simultaneously is denoted by tail dependence. This is part of copula theory as well as multivariate extreme value theory. A profound treatment of copula theory can be found, e.g., in Joe (1997) and Nelsen (2006). Moreover, Mikosch (2003, Chap. 4) gives a nice overview of extreme value theory. The (lower) tail-dependence coefficient of a pair of random variables X and Y or, equivalently, of their joint distribution, is defined as

$$\lambda(X,Y) := \lim_{t \searrow 0} \mathrm{P}\left(F_Y(Y) \le t \mid F_X(X) \le t\right) = \lim_{t \searrow 0} \frac{C(t,t)}{t},$$

[1]The particular choice of the stock indices shall symbolize the nice and fruitful collaboration between Hannu Oja (Finland) and Gabriel Frahm (Germany). Nonetheless, the empirical phenomena that can be observed in the figures occur worldwide for most other stocks and stock indices. The data used in this work have been obtained from VWD (Vereinigte Wirtschaftsdienste GmbH).

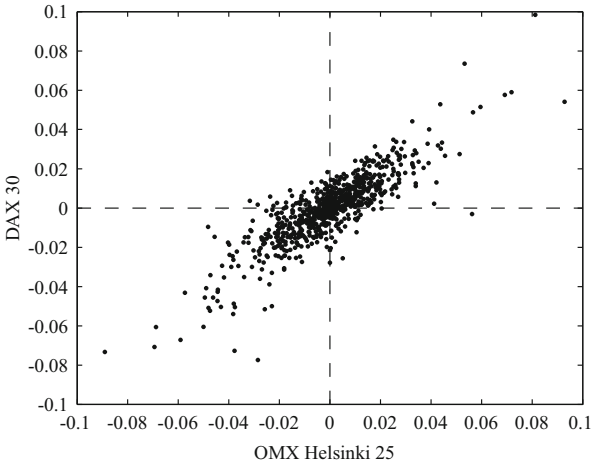

Fig. 17.2 Daily log-returns on OMX Helsinki 25 vs. DAX 30 from 2007-01-03 to 2009-12-31

where C is the copula of (X, Y), F_X is the marginal cumulative distribution function (c.d.f.) of X, and F_Y is the marginal c.d.f. of Y. There exist various ways to extend the concept of tail dependence to the multivariate case (De Luca and Rivieccio 2012; Ferreira and Ferreira 2012; Frahm 2006).

17.2 Elliptical and Generalized Elliptical Distributions

17.2.1 Elliptical Distributions

It is well known that the multivariate normal distribution allows neither for heavy tails nor for tail dependence. To overcome this problem, members of the traditional class of elliptical distributions (Cambanis et al. 1981; Fang et al. 1990; Kelker 1970) are often proposed for the modeling of asset returns (cf., e.g., Bingham and Kiesel 2002; Eberlein and Keller 1995; McNeil et al. 2005, Chap. 3).

In the following, $\mathscr{S}^{k-1} := \left\{\mathbf{u} \in \mathbb{R}^k \colon \|\mathbf{u}\| = 1\right\}$ represents the unit hypersphere, i.e., $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^k$.

Definition 17.1 (Elliptical Distribution) A d-dimensional random vector $\mathbf{X}$ is said to be elliptically distributed if and only if there exist

1. a k-dimensional random vector $\mathbf{U}$, uniformly distributed on $\mathscr{S}^{k-1}$,
2. a nonnegative random variable $\mathscr{R}$ being stochastically independent of $\mathbf{U}$,
3. a vector $\boldsymbol{\mu} \in \mathbb{R}^d$, and a matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times k}$ such that

$$\mathbf{X} = \boldsymbol{\mu} + \mathscr{R}\boldsymbol{\Lambda}\mathbf{U}\,.$$

The random vector $\mathbf{X}$ is said to be spherically distributed if and only if $\mathbf{X} = \mathscr{R}\mathbf{U}$.

We will assume that the location vector $\boldsymbol{\mu}$ is known and so we set $\boldsymbol{\mu} = \mathbf{0}$ without loss of generality. Further, we will call $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$ the dispersion matrix of $\mathbf{X}$ and $\mathscr{R}$ its generating variate.

The main fact that we would like to point out for the further discussion is that elliptical distributions possess two sorts of dependencies, viz.

1. *linear* dependencies, which can be expressed by the dispersion matrix $\boldsymbol{\Sigma}$ and
2. *nonlinear* dependencies imposed by the generating variate $\mathscr{R}$.

For example, consider a bivariate elliptically distributed random vector with components X and Y. Further, suppose that the generating variate $\mathscr{R}$ is regularly varying. This means we have that

$$\mathrm{P}(\mathscr{R} > x) = f(x)\, x^{-\alpha}, \qquad \forall\, x > 0\,,$$

where f is a slowly varying function, i.e., $f(tx)/f(x) \to 1$ as $x \to \infty$ for every $t > 0$. The number $\alpha > 0$ represents the tail index of $\mathscr{R}$ (Mikosch 2003). Thus $\mathrm{P}(\mathscr{R} > x)$ tends to a power law for $x \to \infty$ and $\mathscr{R}$ is said to be "heavy tailed." It is intuitively clear that in this case the two components X and Y are heavy tailed, too. In fact, as is shown by Frahm et al. (2003), the tail-dependence coefficient of X and Y amounts to

$$\lambda = 2\bar{t}_{\alpha+1}\left(\sqrt{\alpha+1}\,\sqrt{\frac{1-\rho}{1+\rho}}\right),$$

where $\bar{t}_\nu$ denotes the survival function of Student's t-distribution with $\nu > 0$ degrees of freedom and ρ is the linear correlation coefficient of X and Y. Hence, the tail dependence is essentially determined by the tail index, α, of $\mathscr{R}$. In particular, the components X and Y can highly depend on each other in a nonlinear way even if they are uncorrelated, i.e., if $\rho = 0$ but $\mathscr{R}$ is regularly varying. The same conclusion can be drawn in the multivariate case (Frahm 2006). Without regular variation, the most evident example, where the components of $\mathbf{X}$ are uncorrelated but (strongly) dependent, is the uniform distribution on a sphere.

17.2.2 Generalized Elliptical Distributions

Elliptical distributions inherit many nice properties from the Gaussian distribution. For example, they are closed under affine-linear transformations, the marginal distributions are also elliptical, and even the conditional distributions remain elliptical. Many elliptical distributions are infinitely divisible, which is an appealing property in the context of financial-data analysis (Bingham and Kiesel 2002). Further, due to the simple stochastic representation of elliptical distributions, they are appropriate

for the modeling of high-dimensional financial data. Nevertheless, they suffer from the property of symmetry. For this reason, we will bear on the class of generalized elliptical distributions (Frahm 2004, Chap. 3).

Definition 17.2 (Generalized Elliptical Distribution) A d-dimensional random vector $\mathbf{X}$ is said to be generalized elliptically distributed if and only if there exist

1. a k-dimensional random vector $\mathbf{U}$, uniformly distributed on $\mathscr{S}^{k-1}$,
2. a random variable $\mathscr{R}$,
3. a vector $\boldsymbol{\mu} \in \mathbb{R}^d$, and a matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times k}$ such that

$$\mathbf{X} = \boldsymbol{\mu} + \mathscr{R}\boldsymbol{\Lambda}\mathbf{U}\,.$$

The random vector $\mathbf{X}$ is said to be generalized spherically distributed if and only if $\mathbf{X} = \mathscr{R}\mathbf{U}$.

All components of elliptical distributions, i.e., the location vector $\boldsymbol{\mu}$, the linear operator $\boldsymbol{\Lambda}$, and the generating variate $\mathscr{R}$, are preserved in Definition 17.2. The only difference is that $\mathscr{R}$ can be negative and even more it may depend on $\mathbf{U}$. This means the radial part of $\mathbf{X}$ may depend on its angular part. This allows us to control for tail dependence and asymmetry. A more detailed discussion regarding the practical implementation of generalized elliptical distributions can be found in Frahm (2004, Sect. 3.4) and Kring et al. (2009).

It is worth pointing out that the class of generalized elliptical distributions does not only include the traditional class of elliptical distributions, but also the class of skew-elliptical distributions (Branco and Dey 2001; Liu and Dey 2004). The latter can be obtained by a modeling technique called hidden truncation (Arnold and Beaver 2004; Frahm 2004, p. 47). However, skew-elliptical distributions have been introduced especially for the modeling of skewness and heavy tails rather than tail dependence (Branco and Dey 2001).

By fitting Student's t-distribution to daily log-returns on stocks, several authors come to the conclusion that the number of degrees of freedom, ν, typically lies between 3 and 7 (see, e.g., McNeil et al. 2005, p. 85). Hence, the t-distribution seems to provide a fairly good fit to daily log-returns. Figure 17.3 contains simulated generalized elliptically distributed daily log-returns. The simulation is based on the idea that ν depends on the direction of the data. More precisely, the log-returns have been simulated as follows:

1. We calculated the eigenvector, $\mathbf{v}$, associated with the larger eigenvalue of the sample covariance matrix of the $n = 756$ daily log-returns depicted in Fig. 17.2. This means we applied a principal-component analysis.
2. Then we simulated $n = 756$ i.i.d. random vectors $\mathbf{X}_t = \mathscr{R}_t\boldsymbol{\Lambda}\mathbf{U}_t$ $(t = 1, 2, \ldots, n)$, where $\boldsymbol{\Lambda}$ denotes the lower triangular Cholesky root of the sample covariance

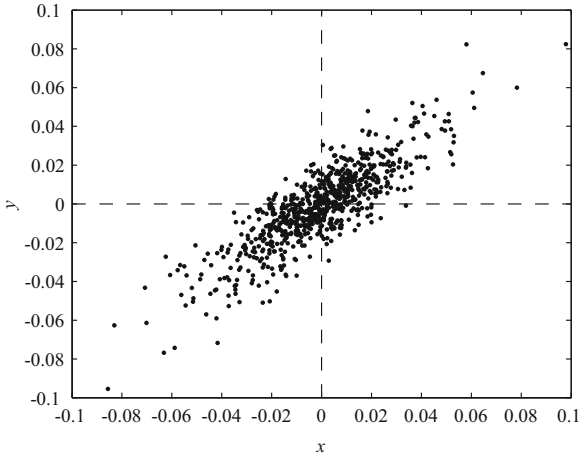

Fig. 17.3 Simulated generalized elliptically distributed daily log-returns ($n = 756$)

matrix.[2] Moreover, the generating variate is given by

$$\mathscr{R}_t = \sqrt{\frac{\chi^2_{t,2}}{\chi^2_{t,\nu}/\nu}}$$

with $\nu = 5 + 95\big(\min\{\angle(\boldsymbol{\Lambda}\mathbf{U}_t, \mathbf{v}), \angle(\boldsymbol{\Lambda}\mathbf{U}_t, -\mathbf{v})\}/(\pi/2)\big)^2$, where $\angle(\mathbf{a}, \mathbf{b})$ denotes the angle between $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$ and $\chi^2_{t,2}$ is independent of $\chi^2_{t,\nu}$ and $\mathbf{U}_t$.

Hence, if $\mathbf{X}$ tends to its first principal component (or to the opposite direction), $\mathscr{R}_t$ has a tail index of $\nu = 5$ and can be considered heavy tailed. By contrast, if it tends to its second principal component, the tail index amounts to $\nu = 100$ and so $\mathscr{R}_t$ is close to the generating variate of a normal distribution. This demonstrates that the class of generalized elliptical distributions is able to reproduce the aforementioned observations regarding the daily log-returns on the stock indices during the financial crisis 2007–2009 (cf. Fig. 17.2).

In virtue of the previous findings, our preliminary conclusions are as follows:

1. The class of generalized elliptical distributions is sufficiently rich. In particular, it includes the class of elliptical and skew-elliptical distributions.
2. The stylized facts of empirical finance can be reproduced by the class of generalized elliptical distributions.
3. This class of distributions seems to be an appropriate model for financial data when investigating standard methods of random matrix theory (RMT).

[2]The sample means of the daily log-returns on the OMX Helsinki 25 and DAX 30 are $-4.7489 \cdot 10^{-4}$ and $-1.3487 \cdot 10^{-4}$, respectively. For this reason, we can simply ignore the location.

The problem is that there exists a tremendous amount of generalized elliptical distribution families that could be considered for the modeling of financial data. Later on we will see that the results given by standard methods of RMT heavily depend on the underlying assumptions concerning the dependence structure of the data and this is essentially determined by the generating variate $\mathscr{R}$. Thus we aim at finding a *distribution-free* approach such that standard methods of RMT can be applied irrespective of the generating variate $\mathscr{R}$.

17.3 Random Matrix Theory

RMT has its origin in nuclear physics, where it has been developed for the modeling of the energy levels of complex nuclei. A contemporary overview of RMT can be found, e.g., in Bai and Silverstein (2010) and Debashis and Aue (2014). During the last years, this topic becomes increasingly important in statistics, particularly in financial-data analysis. For example, Bai et al. (2009), Glombek (2012) and Karoui (2010, 2013) investigate problems of Markowitz portfolio optimization. Moreover, Bouchaud et al. (2003), Laloux et al. (1999) and Plerou et al. (1999, 2002) discuss the application of RMT in the context of principal-component analysis, whereas Bai (2003) and Bai and Ng (2002, 2007) refer to factor analysis.

The spectral distribution of a random matrix $\mathbf{M}$ is defined as follows.

Definition 17.3 (Spectral Distribution) Let $\mathbf{M}$ be a $d \times d$ symmetric random matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$. Then the function

$$F_{\mathbf{M}}(\lambda) = \frac{1}{d} \sum_{i=1}^{d} \mathbf{1}_{\lambda_i \leq \lambda}\,, \qquad \forall\, \lambda \in \mathbb{R}\,,$$

is called the spectral distribution of $\mathbf{M}$.

In multivariate analysis it is usually assumed that the number of dimensions, i.e., d, is fixed. By contrast, in RMT we have that $d \to \infty$ as $n \to \infty$. This makes it possible to derive asymptotic results for high-dimensional data.

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a sequence of d-dimensional random vectors with zero mean. More precisely, we have that $\mathbf{X}_t = (X_{1t}, X_{2t}, \ldots, X_{dt})$ with $\mathrm{E}(X_{it}) = 0$ for each sample element X_{it} $(i, t = 1, 2, \ldots, d, n)$. The sample covariance matrix is given by

$$\mathbf{S} = \frac{1}{n} \sum_{t=1}^{n} \mathbf{X}_t \mathbf{X}_t'\,.$$

In the following, $\mathbf{I}_d$ denotes the $d \times d$ identity matrix and x^+ represents the positive part of $x \in \mathbb{R}$, i.e., $x^+ = \max\{0, x\}$. In RMT it is typically assumed that the sample elements are independent and identically distributed (i.i.d.).

Theorem 17.1 (Bai and Yin 1988) *Suppose that the sample elements are i.i.d., have zero mean, unit variance, and finite fourth moment. Consider the random matrix* $\mathbf{M} = \sqrt{n/d}\,(\mathbf{S} - \mathbf{I}_d)$. *Then for all* $\lambda \in \mathbb{R}$ *the spectral distribution* $F_{\mathbf{M}}(\lambda)$ *converges almost surely to* $F_{\mathrm{W}}(\lambda) = \int_{-\infty}^{\lambda} f_{\mathrm{W}}(x)\,dx$ *with*

$$f_{\mathrm{W}}(x) = \frac{1}{2\pi}\left(4 - x^2\right)^+$$

as $n, d \to \infty$ *and* $n/d \to \infty$.

This theorem guarantees that the eigenspectrum of the sample covariance matrix of a sequence of i.i.d. data converges to Wigner's semicircle law (Wigner 1955, 1958) as $n, d \to \infty$ and $n/d \to \infty$.[3] Indeed, this is a remarkable result, but in many practical applications, the number of observations, n, is not large enough compared to the number of dimensions, d. The following theorem only requires that $n/d \to q$ with $q \in]0, \infty[$ and so the effective sample size n/d can be a small number.

Theorem 17.2 (Bai and Silverstein 2010) *Suppose that the sample elements are i.i.d., have zero mean, and unit variance. Then for all* $\lambda \in \mathbb{R}$ *the spectral distribution* $F_{\mathbf{S}}(\lambda)$ *converges almost surely to* $F_{\mathrm{MP}}(\lambda) = \int_{-\infty}^{\lambda} f_{\mathrm{MP}}(x)\,dx$ *with*

$$f_{\mathrm{MP}}(x) = \frac{q}{2\pi} \cdot \frac{\sqrt{(\lambda_+ - x)^+ (x - \lambda_-)^+}}{x}, \qquad \lambda_\pm = \left(1 \pm \frac{1}{\sqrt{q}}\right)^2,$$

as $n, d \to \infty$ *and* $n/d \to q$ *with* $1 \le q < \infty$. *In case* $0 < q < 1$ *the limiting density is a mixture of a point mass at* 0 *and* $f_{\mathrm{MP}}(x)$ *with weights* $1 - q$ *and* q, *respectively.*

The limiting distribution that is given by Theorem 17.2 is known as the Marčenko–Pastur law (Marčenko and Pastur 1967). It implies that all eigenvalues outside its support $\left[(1 - 1/\sqrt{q})^2, (1 + 1/\sqrt{q})^2\right]$ vanish asymptotically.

17.4 Pitfall and Alternative

17.4.1 Sample Covariance Matrix

Consider a sample of 500-dimensional random vectors with sample size $n = 1000$, where the vector components are mutually independent and possess a standardized univariate t-distribution with 5 degrees of freedom. In the subsequent discussion this is said to be a multivariate *non-elliptical* t-distribution. The left-hand side of Fig. 17.4 contains the eigenspectrum obtained by the sample covariance matrix.

[3] The semicircle law implies that all eigenvalues outside its support $[-2, 2]$ vanish asymptotically.

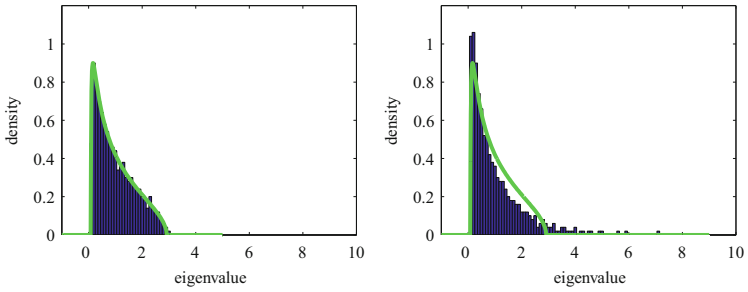

Fig. 17.4 Eigenspectra of **S** based on non-elliptically, i.e., independent (*left*), vs. elliptically, i.e., uncorrelated (*right*), multivariate t-distributed data ($n = 1000, d = 500$) with 5 degrees of freedom. The *green curve* represents the density function of the Marčenko–Pastur law for $q = 2$

Obviously, this is consistent with the Marčenko–Pastur law. By contrast, suppose that the random vectors have a standardized multivariate *elliptical* t-distribution with 5 degrees of freedom. More precisely, it is supposed that the vector components are uncorrelated, i.e., $\boldsymbol{\Sigma} \propto \mathbf{I}_{500}$, but not independent. In this case the Marčenko–Pastur law is clearly violated.

More precisely, we find 29 spurious eigenvalues exceeding the Marčenko–Pastur upper bound $\lambda_+ = (1 + 1/\sqrt{2})^2 = 2.91$ and the largest eigenvalue corresponds to 7.0698. In the physics literature, the exceeding eigenvalues are often considered "signals" or "information" (see, e.g., Bouchaud et al. 2003; Laloux et al. 1999; Plerou et al. 1999, 2002). In terms of principal-component analysis, the exceeding eigenvalues could be interpreted as the contribution of the first principal components to the total variation of the data. Figure 17.4 demonstrates that in this case we would seriously overestimate the systematic risk of asset returns.

Theorem 17.2 does not require any specific distributional assumption. In the context of elliptical and generalized elliptical distributions this is a potential fallacy. It is well known that the multivariate normal is the only elliptical distribution that allows for independent components. Hence, the problem is that the components of a spherically distributed random vector $\mathbf{X} = \mathscr{R}\mathbf{U}$ are not independent unless $\mathscr{R} \propto \chi_d$, i.e., if $\mathbf{X}$ has a multivariate normal distribution. For example, if asset returns follow a multivariate elliptical t-distribution, they might be uncorrelated but never independent in the cross section. More precisely, short-term asset returns are tail dependent. This is the reason why the Marčenko–Pastur law in general does not work for spherically distributed data.[4] We often have observations suggesting that the vector components are highly correlated. The smaller the tail index of the generating variate $\mathscr{R}$, i.e., the heavier the tails of $\mathbf{X}$, the more spurious eigenvalues occur.

[4] We know only two exceptions, i.e., (1) $\mathscr{R} = \chi_d$ and (2) $\mathscr{R} = \sqrt{d}$ (Marčenko and Pastur 1967).

17.4.2 Tyler's M-Estimator

Since daily asset returns follow a leptokurtic or heavy-tailed distribution, it seems natural to use a robust covariance matrix as an alternative to the sample covariance matrix. In the following discussion we focus on Tyler's M-estimator (Tyler 1987a). Its many nice properties have been established, e.g., by Adrover (1998), Dümbgen (1998), Dümbgen and Tyler (2005), Frahm (2009), Frahm and Glombek (2012), Kent and Tyler (1988), Kent and Tyler (1991), Maronna and Yohai (1990), Paindaveine (2008), and Tyler (1987b), etc. We do not take other estimators into consideration, because Tyler's M-estimator turns out to be a canonical choice in the context of financial time series. This will become clear by the end of this section.

If the log-returns are elliptically distributed and the second moment of $\mathscr{R}$ is finite, we have that $\mathrm{Var}(\mathbf{X}) = \mathrm{E}(\mathscr{R}^2)/k \cdot \boldsymbol{\Sigma}$. However, in many applications of multivariate data analysis we need to know only the shape matrix of $\mathbf{X}$, i.e., $\boldsymbol{\Omega} = \boldsymbol{\Sigma}/\sigma^2(\boldsymbol{\Sigma})$, where σ^2 is any scale function, i.e., a positive homogeneous function of degree 1 such that $\sigma^2(\mathbf{I}_d) = 1$.[5] The shape matrix $\boldsymbol{\Omega}$ reflects the linear dependence structure of $\mathbf{X}$. Since the covariance matrix of $\mathbf{X}$ is proportional to $\boldsymbol{\Sigma}$, $\mathbf{S}/\sigma^2(\mathbf{S})$ represents a consistent estimator for $\boldsymbol{\Omega}$. In general, this is not satisfied if $\mathscr{R}$ depends on $\mathbf{U}$, i.e., if $\mathbf{X}$ is not elliptically distributed. In the subsequent discussion it is shown that Tyler's M-estimator is a canonical estimator for the linear dependence structure of any generalized elliptically distributed random vector $\mathbf{X}$.

We assume that $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Lambda} \in \mathbb{R}^{d\times k}$ with $\mathrm{rk}\,\boldsymbol{\Lambda} = d$, and $\mathrm{P}(\mathscr{R} = 0) = 0$, i.e., $\mathbf{X}$ has no point mass at its origin. Due to the stochastic representation of $\mathbf{X}$ given by Definition 17.2, the following relations hold:

$$\frac{\mathbf{X}}{\|\mathbf{X}\|} = \frac{\mathscr{R}\boldsymbol{\Lambda}\mathbf{U}}{\|\mathscr{R}\boldsymbol{\Lambda}\mathbf{U}\|} = \mathrm{sign}(\mathscr{R})\frac{\boldsymbol{\Lambda}\mathbf{U}}{\|\boldsymbol{\Lambda}\mathbf{U}\|} = \mathrm{sign}(\mathscr{R})\mathbf{V}\,, \qquad \mathbf{V} := \boldsymbol{\Lambda}\mathbf{U}\,. \tag{17.1}$$

The unit random vector $\mathrm{sign}(\mathscr{R})\mathbf{V}$ does not depend on the absolute value of $\mathscr{R}$. In particular, it is invariant under the occurrence of extreme values of $\mathscr{R}$. Nonetheless, $\mathrm{sign}(\mathscr{R})$ cannot be cancelled out and indeed it may depend on $\mathbf{U}$.

Suppose for the moment that $\mathrm{sign}(\mathscr{R})$ was known for each realization of $\mathscr{R}$, so that we can easily calculate every realization of $\mathbf{V}$, i.e., $\mathbf{V}_t = \mathrm{sign}(\mathscr{R}_t)\mathbf{X}_t/\|\mathbf{X}_t\|$ for $t = 1, 2, \ldots, n$. The distribution of $\mathbf{V}$ depends on $\boldsymbol{\Lambda}$ only through $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$ and thus we can estimate $\boldsymbol{\Sigma}$ by maximum likelihood. Interestingly, for this purpose we need no assumption about the generating variate $\mathscr{R}$. Even the dependence structure of $\mathscr{R}$ and $\mathbf{U}$ is not relevant. Hence, the resulting estimator is distribution-free.

For deriving the corresponding ML-estimator, we have to calculate the density function of $\mathbf{V}$ and search for $\mathbf{T} = \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^\top$ with

[5] For example, we could choose $\sigma^2(\boldsymbol{\Sigma}) = (\mathrm{tr}\,\boldsymbol{\Sigma})/d$ so that $\mathrm{tr}\,\boldsymbol{\Omega} = d$ or $\sigma^2(\boldsymbol{\Sigma}) = (\det\boldsymbol{\Sigma})^{1/d}$ so that $\det\boldsymbol{\Sigma} = 1$ (Frahm 2009; Paindaveine 2008).

$$\boldsymbol{\Upsilon} = \arg\max_{\boldsymbol{\Lambda}} \prod_{t=1}^{n} \psi\left(\mathbf{V}_t ; \boldsymbol{\Lambda}\right),$$

where $\psi(\mathbf{v})$ represents the density of $\mathbf{V}$ at $\mathbf{v} \in \mathscr{S}^{d-1}$. In the following theorem it is assumed without loss of generality that $\det \boldsymbol{\Lambda} = \det \boldsymbol{\Sigma} = 1$.[6]

Theorem 17.3 *Let $\boldsymbol{\Lambda}$ be a $d \times k$ matrix with $\operatorname{rk} \boldsymbol{\Lambda} = d$ and $\det \boldsymbol{\Lambda} = 1$. Further, consider the matrix $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\top}$. The density of the unit random vector $\mathbf{V} = \boldsymbol{\Lambda}\mathbf{U}/\|\boldsymbol{\Lambda}\mathbf{U}\|$ with respect to the uniform measure on $\mathscr{S}^{d-1}$ corresponds to*

$$\psi(\mathbf{v}) = \frac{\boldsymbol{\Gamma}(d/2)}{2\pi^{d/2}} \cdot \sqrt{\mathbf{v}^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{v}}^{-d}$$

for all $\mathbf{v} \in \mathscr{S}^{d-1}$.

Proof See, e.g., Frahm (2004, pp. 59–60).

The distribution given by Theorem 17.3 is the angular central Gaussian distribution (Tyler 1987b; Watson 1983). Due to the Courant-Fischer Theorem, the density function ψ has a local extremum at $\mathbf{w} \in \mathscr{S}^{d-1}$ if and only if $\mathbf{w}$ is an eigenvector of $\boldsymbol{\Sigma}$ and we have that $\psi(\mathbf{w}) \propto \lambda^{d/2}$, where λ is the eigenvalue associated with $\mathbf{w}$.

Figure 17.5 exemplifies the density function of the angular central Gaussian. Note that ψ is symmetric, i.e., $\psi(\mathbf{v}) = \psi(-\mathbf{v})$, and thus it is not necessary to know the sign of $\mathscr{R}$ for calculating the ML-estimator based on the density function of the angular central Gaussian. This means our previous assumption is superfluous.

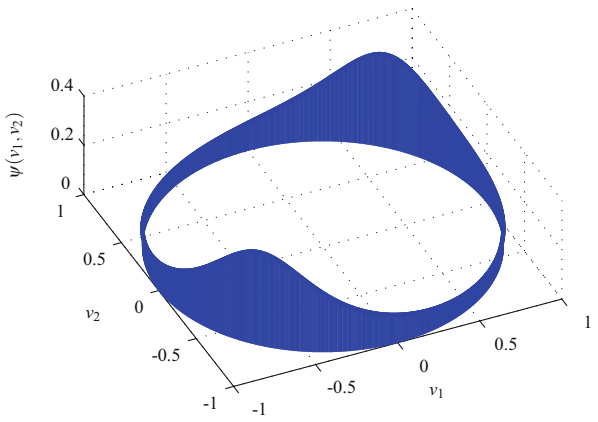

Fig. 17.5 Density of the angular central Gaussian distribution of a two-dimensional unit random vector generated by $\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22} \propto 1$ and $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21} \propto 0.7$

[6] This sort of normalization can be considered canonical (Paindaveine 2008).

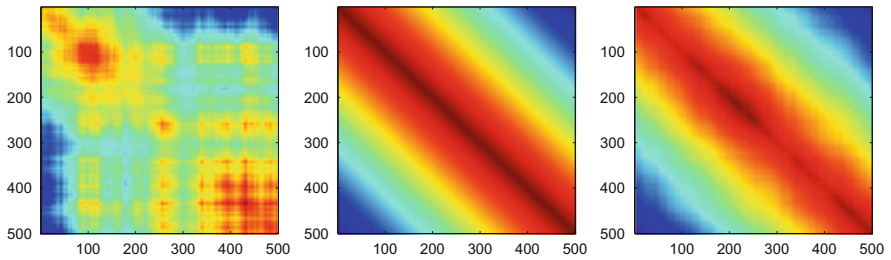

Fig. 17.6 True dispersion matrix (*middle*), sample covariance matrix (*left*), and Tyler's M-estimator (*right*). The estimates are based on a sample of multivariate elliptically t-distributed observations with sample size $n = 1000$, $d = 500$ dimensions, and $\nu = 2$ degrees of freedom

Now, consider a sample of generalized elliptically distributed observations $\mathbf{X}_1$, $\mathbf{X}_2, \ldots, \mathbf{X}_n$. As noted by Tyler (1987b) and Frahm (2004, Sect. 4.2.2),[7] the desired ML-estimator is given by the fixed-point equation

$$\mathbf{T} = \frac{d}{n} \sum_{t=1}^{n} \frac{\mathbf{V}_t \mathbf{V}_t'}{\mathbf{V}_t' \mathbf{T}^{-1} \mathbf{V}_t} \,. \tag{17.2}$$

Actually, this corresponds to Tyler's M-estimator (Tyler 1987a), i.e.,

$$\mathbf{T} = \frac{d}{n} \sum_{t=1}^{n} \frac{\mathbf{X}_t \mathbf{X}_t^\top}{\mathbf{X}_t^\top \mathbf{T}^{-1} \mathbf{X}_t} \,.$$

If the solution of this fixed-point equation exists, it is unique only up to a scaling constant. This means $\mathbf{T}$ must be normalized in any way and in the following we assume that $\operatorname{tr} \mathbf{T} = d$.

The right-hand side of Fig. 17.6 contains a realization of Tyler's M-estimator $\mathbf{T}$, based on a simulated sample of $n = 1000$ multivariate elliptically t-distributed observations with $d = 500$ dimensions and $\nu = 2$ degrees of freedom. The true dispersion matrix $\boldsymbol{\Sigma}$ is a symmetric Toeplitz matrix, which can be seen in the middle of Fig. 17.6. The corresponding realization of the sample covariance matrix $\mathbf{S}$ is depicted on the left-hand side of Fig. 17.6. Obviously, $\mathbf{T}$ is a robust alternative to $\mathbf{S}$.

At the beginning of this section we claimed that $\mathbf{T}$ is a canonical choice when dealing with financial time-series data. Asset returns typically exhibit nonlinear dependencies both in the cross section and in time. We already mentioned that the tail-dependence coefficient of an elliptically distributed random vector $\mathbf{X}$ essentially depends on the tail index of its generating variate. Now, suppose that the time series $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ is such that $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n$ are serially independent, but in contrast

[7] Tyler (1987b) refers only to elliptical distributions, whereas Frahm (2004) observes that the same result applies as well to generalized elliptical distributions.

$\mathscr{R}_1, \mathscr{R}_2, \ldots, \mathscr{R}_n$ have a serial dependence structure. For example, the log-returns could be conditionally heteroscedastic. Our key note is that $\mathbf{T}$ depends only on $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n$, i.e., the angular part, but not on $\mathscr{R}_1, \mathscr{R}_2, \ldots, \mathscr{R}_n$, i.e., the radial part of the data. This can be seen by re-writing Eq. (17.2):

$$\mathbf{T} = \frac{d}{n} \sum_{t=1}^{n} \frac{\boldsymbol{\Lambda} \mathbf{U}_t \mathbf{U}_t^\top \boldsymbol{\Lambda}^\top}{\mathbf{U}_t^\top \boldsymbol{\Lambda}^\top \mathbf{T}^{-1} \boldsymbol{\Lambda} \mathbf{U}_t} .$$

Hence, the solution of the fixed-point equation does not depend on $\mathscr{R}_1, \mathscr{R}_2, \ldots, \mathscr{R}_n$ and so it does not matter how the sequence $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ is driven by the generating variates both in the cross section or in time. This means the asset returns might depend on each other through their generating variates in an *arbitrary* way. Even the finite-sample distribution of $\mathbf{T}$ is not influenced by $\mathscr{R}_1, \mathscr{R}_2, \ldots, \mathscr{R}_n$. This makes $\mathbf{T}$ highly favorable for heavy-tailed financial time series, irrespective of whether the sample size, n, is large or small or the number of dimensions, d, is high or low.

17.5 Spectral Properties of Tyler's M-Estimator

In virtue of the aforementioned results we can expect that $\mathbf{T}$ is an appropriate alternative to $\mathbf{S}$ in the context of RMT. This is confirmed by the next theorem.

Theorem 17.4 (Frahm and Glombek 2012) *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a sequence of d-dimensional generalized spherically distributed random vectors whose angular parts $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n$ are mutually independent. Consider the random matrix $\mathbf{M} = \sqrt{n/d}\,(\mathbf{T} - \mathbf{I}_d)$ with $\operatorname{tr} \mathbf{T} = d$. Then for all $\lambda \in \mathbb{R}$ the spectral distribution $F_{\mathbf{M}}(\lambda)$ converges in probability to $F_{\mathrm{W}}(\lambda) = \int_{-\infty}^{\lambda} f_{\mathrm{W}}(x)\,dx$ with*

$$f_{\mathrm{W}}(x) = \frac{1}{2\pi}\left(4 - x^2\right)^+$$

as $n, d \to \infty$ and $n/d \to \infty$.

Hence, after an appropriate normalization, the spectral distribution of Tyler's M-estimator converges in probability to Wigner's semicircle law as $n, d \to \infty$ but $n/d \to \infty$.[8] Hence, in contrast to Theorem 17.1, the components of $\mathbf{X}$ are not required to be independent. For other results related to Tyler's M-estimator in the case $n, d \to \infty$ and $n/d \to \infty$ see Dümbgen (1998).

[8] In contrast to Theorems 17.1 and 17.2, Theorem 17.4 only states that the spectral distribution converges in probability but not almost surely. More details on the technical difficulties related to the proof of strong consistency can be found at the end of Frahm and Glombek (2012).

The remaining question is whether the spectral distribution of $\mathbf{T}$ converges to the Marčenko–Pastur law in case $n/d \to q < \infty$. This is formalized by the following conjecture.

Conjecture *Suppose that one of the following conditions is satisfied:*

1. *The sample elements are i.i.d., have zero mean, finite variance, and a continuous distribution.*
2. $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ *is a sequence of d-dimensional generalized spherically distributed random vectors whose angular parts* $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n$ *are mutually independent.*

Then for all $\lambda \in \mathbb{R}$ *the spectral distribution* $F_{\mathbf{T}}(\lambda)$ *converges in probability to* $F_{\mathrm{MP}}(\lambda) = \int_{-\infty}^{\lambda} f_{\mathrm{MP}}(x)\,dx$ *with*

$$f_{\mathrm{MP}}(x) = \frac{q}{2\pi} \cdot \frac{\sqrt{(\lambda_+ - x)^+ (x - \lambda_-)^+}}{x}, \qquad \lambda_\pm = \left(1 \pm \frac{1}{\sqrt{q}}\right)^2,$$

as $n, d \to \infty$ *and* $n/d \to q$ *with* $1 \le q < \infty$. *In case* $0 < q < 1$ *the limiting density is a mixture of a point mass at* 0 *and* $f_{\mathrm{MP}}(x)$ *with weights* $1 - q$ *and* q, *respectively.*

Figure 17.7 demonstrates our conjecture. In contrast to Fig. 17.6 we can see that the spectral distribution of $\mathbf{T}$ converges to the Marčenko–Pastur law both if the data are independent and if they are only uncorrelated.

Our arguments can be understood as follows. Suppose that the first condition of our conjecture is satisfied. Then we have that $\mathrm{E}(\mathbf{X}\mathbf{X}^\top) = \sigma^2 \mathbf{I}_d$, where $\sigma^2 > 0$ is the variance of any sample element. Suppose that d is large. From the Law of Large Numbers we conclude that $\mathbf{X}^\top\mathbf{X}/d \approx \sigma^2$ and thus $\mathrm{E}\big[(\mathbf{X}\mathbf{X}^\top)/(\mathbf{X}^\top\mathbf{X}/d)\big] \approx \mathbf{I}_d$. For this reason, Tyler's M-estimator

$$\mathbf{T} = \frac{1}{n} \sum_{t=1}^{n} \frac{\mathbf{X}_t \mathbf{X}_t^\top}{\mathbf{X}_t^\top \mathbf{T}^{-1} \mathbf{X}_t / d}$$

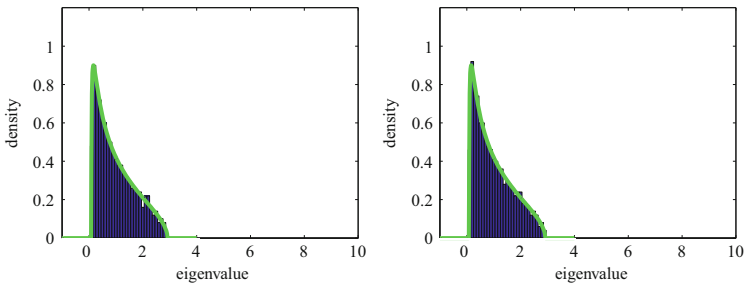

Fig. 17.7 Eigenspectra of $\mathbf{T}$ based on non-elliptically, i.e., independent (*left*), vs. elliptically, i.e., uncorrelated (*right*), multivariate t-distributed data ($n = 1000, d = 500$) with 5 degrees of freedom. The *green curve* represents the density function of the Marčenko–Pastur law for $q = 2$

converges almost surely to a $d \times d$ matrix that is close to $\mathbf{I}_d$ as $n \to \infty$ (Tyler 1987a). This means for all $t = 1, 2, \ldots, n$ we expect that $\mathbf{X}_t^\top \mathbf{T}^{-1} \mathbf{X}_t / d \to_{\text{a.s.}} \sigma^2$ as $n, d \to \infty$ and thus $\mathbf{T} \approx \mathbf{S}$ if d and n are large. Here $\mathbf{S}$ denotes the sample covariance matrix of $\mathbf{X}_1/\sigma, \mathbf{X}_2/\sigma, \ldots, \mathbf{X}_n/\sigma$, which satisfies the global i.i.d. assumption of Theorem 17.2.

By contrast, if the second condition is satisfied, Tyler's M-estimator corresponds to

$$\mathbf{T} = \frac{d}{n} \sum_{t=1}^{n} \frac{\mathbf{U}_t \mathbf{U}_t^\top}{\mathbf{U}_t^\top \mathbf{T}^{-1} \mathbf{U}_t} = \frac{d}{n} \sum_{t=1}^{n} \frac{(\chi_{d,t} \mathbf{U}_t)(\chi_{d,t} \mathbf{U}_t^\top)}{(\chi_{d,t} \mathbf{U}_t^\top) \mathbf{T}^{-1} (\chi_{d,t} \mathbf{U}_t)} = \frac{1}{n} \sum_{t=1}^{n} \frac{\mathbf{Y}_t \mathbf{Y}_t^\top}{\mathbf{Y}_t^\top \mathbf{T}^{-1} \mathbf{Y}_t / d},$$

where $\mathbf{Y}_t = \chi_{d,t} \mathbf{U}_t$ $(t = 1, 2, \ldots, n)$ is a sequence of independent standard normally distributed random vectors. Now, due to the same arguments we obtain $\mathbf{T} \approx \mathbf{S}$, where $\mathbf{S}$ denotes the sample covariance matrix of $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$. It is clear that also the latter sample satisfies the global i.i.d. assumption of Theorem 17.2. Thus our intuition tells us that $F_{\mathbf{T}}$ and $F_{\mathbf{S}}$ converge to the same limit, i.e., to the Marčenko–Pastur law, if $n, d \to \infty$ with $n/d \to q < \infty$. Nonetheless, the proof of our conjecture is formidable and, to the best of our knowledge, still missing in the literature.

Acknowledgements We thank David Tyler for his valuable and encouraging comments as well as Karl Mosler for his important suggestions. We also thank Lutz Dümbgen, Marc Hallin, Frank Hampel, Ingo Klein, Hans-Rudolf Künsch, Alexander McNeil, Davy Paindaveine, and Bernd Rosenow for our fruitful discussions. Special thanks go to Hannu Oja. Without his scientific contributions this paper might have never been accomplished.

References

Adrover, J.G.: Minimax bias-robust estimation of the dispersion matrix of a multivariate distribution. Ann. Stat. **26**, 2301–2320 (1998)

Arnold, B.C., Beaver, R.J.: Elliptical models subject to hidden truncation or selective sampling. In: Genton, M.G. (ed.) Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality, Chapter 6. Chapman & Hall, Boca Raton (2004)

Bai, J.: Inferential theory for factor models for large dimensions. Econometrica **71**, 135–171 (2003)

Bai, J., Ng, S.: Determining the number of factors in approximate factor models. Econometrica **70**, 191–221 (2002)

Bai, J., Ng, S.: Determining the number of primitive shocks in factor models. J. Bus. Econ. Stat. **25**, 52–60 (2007)

Bai, Z.D., Silverstein, J.W.: Spectral Analysis of Large Dimensional Random Matrices, 2nd edn. Springer, Berlin (2010)

Bai, Z.D., Yin, Y.Q.: Convergence to the semicircle law. Ann. Probab. **16**, 863–875 (1988)

Bai, Z.D., Liu, H., Wong, W.K.: On the Markowitz mean-variance analysis of self-financing portfolios. Risk Decision Anal. **1**, 35–42 (2009)

Bingham, N.H., Kiesel, R.: Semi-parametric modelling in finance: theoretical foundation. Quant. Finance **2**, 241–250 (2002)

Bouchaud, J.P., Cont, R., Potters, M.: Scaling in stock market data: stable laws and beyond. In: Dubrulle, B., Graner, F., Sornette, D. (eds.) Scale Invariance and Beyond. Proceedings of the CNRS Workshop on Scale Invariance, Les Houches, March 1997. EDP-Springer, Berlin (1997)

Bouchaud, J.P., Mézard, M., Potters, M.: Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management. Cambridge University Press, Cambridge (2003)

Branco, M.D., Dey, D.K.: A general class of multivariate skew-elliptical distributions. J. Multivar. Anal. **79**, 99–113 (2001)

Breymann, W., Dias, A., Embrechts, P.: Dependence structures for multivariate high-frequency data in finance. Quant. Finance **3**, 1–14 (2003)

Cambanis, S., Huang, S., Simons, G.: On the theory of elliptically contoured distributions. J. Multivar. Anal. **11**, 368–385 (1981)

De Luca, G., Rivieccio, G.: Multivariate tail dependence coefficients for Archimedean copulae. In: Di Ciaccio, A., Coli, M., Ibanez, J.M.A. (eds.) Advanced Statistical Methods for the Analysis of Large Data-Sets. Studies in Theoretical and Applied Statistics, pp. 287–296. Springer, Berlin (2012)

Debashis, P., Aue, A.: Random matrix theory in statistics: a review. J. Stat. Plan. Inference **150**, 1–29 (2014)

Ding, Z., Granger, C.W.J., Engle, R.F.: A long memory property of stock market returns and a new model. J. Empir. Finance **1**, 83–106 (1993)

Dümbgen, L.: On Tyler's M-functional of scatter in high dimension. Ann. Inst. Stat. Math. **50**, 471–491 (1998)

Dümbgen, L., Tyler, D.E.: On the breakdown properties of some multivariate M-functionals. Scand. J. Stat. **32**, 247–264 (2005)

Eberlein, E., Keller, U.: Hyperbolic distributions in finance. Bernoulli **1**, 281–299 (1995)

Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling Extremal Events (for Insurance and Finance). Springer, Berlin (1997)

Engle, R.F.: Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica **50**, 987–1007 (1982)

Fama, E.F.: The behavior of stock market prices. J. Bus. **38**, 34–105 (1965)

Fang, K.T., Kotz, S., Ng, K.W.: Symmetric Multivariate and Related Distributions. Chapman & Hall, Boca Raton (1990)

Ferreira, H., Ferreira, M.: On extremal dependence of block vectors. Kybernetika **48**, 988–1006 (2012)

Frahm, G.: Generalized elliptical distributions: theory and applications. Ph.D. thesis, University of Cologne (2004)

Frahm, G.: On the extremal dependence coefficient of multivariate distributions. Stat. Probab. Lett. **76**, 1470–1481 (2006)

Frahm, G.: Asymptotic distributions of robust shape matrices and scales. J. Multivar. Anal. **100**, 1329–1337 (2009)

Frahm, G., Glombek, K.: Semicircle law of Tyler's M-estimator for scatter. Stat. Probab. Lett. **82**, 959–964 (2012)

Frahm, G., Junker, M., Szimayer, A.: Elliptical copulas: applicability and limitations. Stat. Probab. Lett. **63**, 275–286 (2003)

Glombek, K.: High-dimensionality in statistics and portfolio optimization. Ph.D. thesis, University of Cologne (2012)

Joe, H.: Multivariate Models and Dependence Concepts. Chapman & Hall, Boca Raton (1997)

Junker, M., May, A.: Measurement of aggregate risk with copulas. Econ. J. **8**, 428–454 (2005)

Karoui, N.E.: High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: risk underestimation. Ann. Stat. **38**, 3487–3566 (2010)

Karoui, N.E.: On the realized risk of high-dimensional Markowitz portfolios. SIAM J. Financ. Math. **4**, 737–783 (2013)

Kelker, D.: Distribution theory of spherical distributions and a location-scale parameter generalization. Sankhya A **32**, 419–430 (1970)

Kent, J.T., Tyler, D.E.: Maximum likelihood estimation for the wrapped Cauchy distribution. J. Appl. Stat. **15**, 247–254 (1988)
Kent, J.T., Tyler, D.E.: Redescending M-estimates of multivariate location and scatter. Ann. Stat. **19**, 2102–2119 (1991)
Kring, S., Rachev, S.T., Höchstötter, M.: Multi-tail generalized elliptical distributions for asset returns. Econ. J.**12**, 272–291 (2009)
Laloux, L., Cizeau, P., Bouchaud, J.P.: Noise dressing of financial correlation matrices. Phys. Rev. Lett. **83**, 1467–1470 (1999)
Liu, J., Dey, D.K.: Skew-elliptical distributions. In: Genton, M.G. (ed.) Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality, Chapter 3. Chapman & Hall, Boca Raton (2004)
Mandelbrot, B.: The variation of certain speculative prices. J. Bus. **36**, 394–419 (1963)
Maronna, R., Yohai, V.: The maximum bias of robust covariances. Commun. Stat. Theory Methods **19**, 3925–3933 (1990)
Marčenko, V.A., Pastur, L.A.: Distribution of eigenvalues for some sets of random matrices. Math. USSR Sb. **72**, 457–483 (1967)
McNeil, A.J., Frey, R., Embrechts, P.: Quantitative Risk Management. Princeton University Press, Princeton (2005)
Mikosch, T.: Modeling dependence and tails of financial time series. In: Finkenstaedt, B., Rootzén, H., (eds.) Extreme Values in Finance, Telecommunications, and the Environment. Chapman & Hall, Boca Raton (2003)
Nelsen, R.B.: An Introduction to Copulas, 2nd edn. Springer, Berlin (2006)
Paindaveine, D.: A canonical definition of shape. Stat. Probab. Lett. **78**, 2240–2247 (2008)
Plerou, V., Gopikrishnan, P., Rosenow, B.: Universal and nonuniversal properties of cross correlations in financial time series. Phys. Rev. Lett. **83**, 1471–1474 (1999)
Plerou, V., Gopikrishnan, P., Rosenow, B.: Random matrix approach to cross correlations in financial data. Phys. Rev. E **65**, Art. no. 066126 (2002)
Tyler, D.E.: A distribution-free M-estimator of multivariate scatter. Ann. Stat. **15**, 234–251 (1987a)
Tyler, D.E.: Statistical analysis for the angular central Gaussian distribution on the sphere. Biometrika **74**, 579–589 (1987b)
Watson, G.S.: Statistics on Spheres. Wiley, New York (1983)
Wigner, E.P.: Characteristic vectors of bordered matrices with infinite dimensions. Ann. Math. **62**, 548–564 (1955)
Wigner, E.P.: On the distributions of the roots of certain symmetric matrices. Ann. Math. **67**, 325–327 (1958)

Chapter 18
Affine Equivariant Rank-Weighted L-Estimation of Multivariate Location

Pranab Kumar Sen, Jana Jurečková, and Jan Picek

Abstract In the multivariate one-sample location model, only L-estimators are affine equivariant; we propose a class of flexible robust, affine equivariant L-estimators of location, for distributions invoking affine-invariance of Mahalanobis distances of individual observations. An involved iteration process for their computation is numerically illustrated.

Keywords Affine equivariance • Affine invariance • D-optimality • Mahalanobis distance and ordering • Nearest neighbors • Outlyingness

18.1 Introduction

The affine-equivariance and its dual affine-invariance are natural generalizations of univariate translation-scale equivariance and invariance notions (Eaton 1983). Consider the group $\mathscr{C}$ of transformations of $\mathbb{R}_p$ to $\mathbb{R}_p$:

$$\mathbf{X} \mapsto \mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{b},\ \mathbf{b} \in \mathbb{R}_p \tag{18.1}$$

where $\mathbf{B}$ is a positive definite $p \times p$ matrix. Generally with the choice of $\mathbf{B}$ we do not transform dependent coordinates of $\mathbf{X}$ to stochastically independent coordinates of $\mathbf{Y}$. This is possible when $\mathbf{X}$ has a multinormal distribution with mean vector $\boldsymbol{\theta}$ and positive definite dispersion matrix $\boldsymbol{\Sigma}$, when letting $\boldsymbol{\Sigma}^{-1} = \mathbf{B}\mathbf{B}^\top$, so that

P.K. Sen (✉)
Departments of Statistics and Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
e-mail: pksen@bios.unc.edu

J. Jurečková
Faculty of Mathematics and Physics, Department of Probability and Statistics, Charles University, CZ-186 75, Prague 8, Czech Republic
e-mail: jurecko@karlin.mff.cuni.cz

J. Picek
Department of Applied Mathematics, Technical University of Liberec, Liberec, Czech Republic
e-mail: jan.picek@tul.cz

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_18

$\mathbb{E}\mathbf{Y} = \boldsymbol{\xi} = \mathbf{B}\boldsymbol{\theta} + \mathbf{b}$ and dispersion matrix $\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top = \mathbf{I}_p$. To construct and study the affine equivariant estimator of the location $\boldsymbol{\theta}$ we need to consider some affine-invariant (AI) norm. The most well-known affine invariant norm is the Mahalanobis norm, whose squared version is

$$\boldsymbol{\Delta}^2 = (\mathbf{X} - \boldsymbol{\theta})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\theta}) = \|\mathbf{X} - \boldsymbol{\theta}\|_\Sigma^2 \tag{18.2}$$

where $\boldsymbol{\Sigma}$ is the dispersion matrix of $\mathbf{X}$. To incorporate this norm, we need to use its empirical version based on independent sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$, namely

$$s_{ij} = \frac{1}{2}(\mathbf{X}_i - \mathbf{X}_j)^\top \mathbf{V}_n^{*-1}(\mathbf{X}_i - \mathbf{X}_j),\ 1 \le i < j \le n$$

where $\mathbf{V}_n^* = (n(n-1))^{-1} \sum_{1 \le i < j \le n} (\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^\top$. To avoid redundancy, we may consider the reduced set

$$\tilde{d}_{ni} = (\mathbf{X}_i - \mathbf{X}_n)^\top \tilde{\mathbf{V}}_n^{-1}(\mathbf{X}_i - \mathbf{X}_n),\ i = 1, \ldots, n-1 \tag{18.3}$$

$$\tilde{\mathbf{V}}_n = \sum_{i=1}^{n-1} (\mathbf{X}_i - \mathbf{X}_n)(\mathbf{X}_i - \mathbf{X}_n)^\top \tag{18.4}$$

which forms the maximal invariant (MI) with respect to affine transformations (18.1). An equivalent form of the maximal invariant is

$$d_{ni} = (\mathbf{X}_i - \overline{\mathbf{X}}_n)^\top \mathbf{V}_n^{-1}(\mathbf{X}_i - \overline{\mathbf{X}}_n),\ i = 1, \ldots, n \tag{18.5}$$

$$\mathbf{V}_n = \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}}_n)(\mathbf{X}_i - \overline{\mathbf{X}}_n)^\top \tag{18.6}$$

(Obenchain 1971). Note that all the d_{ni} are between 0 and 1 and their sum equals to p. Because d_{ni} are exchangeable, bounded random variables, all nonnegative, with a constant sum equal to p, the asymptotic properties of the array $(d_{n1}, \ldots, d_{nn})^\top$ follow from Chernoff and Teicher (1958) and Weber (1980). Similarly, $\sum_{i=1}^{n-1} \tilde{d}_{ni} = p$. Neither $\overline{\mathbf{X}}_n$ nor $\mathbf{V}_n$ is robust against outliers and gross errors contamination. As such, we are motivated to replace $\overline{\mathbf{X}}_n$ and $\mathbf{V}_n$ by suitable robust versions and incorporate them in the formulation of a robust affine equivariant estimator of $\boldsymbol{\theta}$. If $\hat{\boldsymbol{\theta}}_n$ is some affine equivariant estimator of $\boldsymbol{\theta}$, then writing

$$\hat{\mathbf{V}}_n = \sum_{i=1}^{n} (\mathbf{X}_i - \hat{\boldsymbol{\theta}}_n)(\mathbf{X}_i - \hat{\boldsymbol{\theta}}_n)^\top$$

we may note that $\hat{\mathbf{V}}_n$ is smaller than $\mathbf{V}_n$ in the matrix sense. However, it cannot be claimed that the Mahalanobis distances (18.5) can be made shorter by using $\hat{\boldsymbol{\theta}}_n$ instead of $\overline{\mathbf{X}}_n$, because $\sum_{i=1}^n (\mathbf{X}_i - \hat{\boldsymbol{\theta}}_n)^\top \hat{\mathbf{V}}_n^{-1} (\mathbf{X}_i - \hat{\boldsymbol{\theta}}_n) = p$. Our motivation is to employ the robust Mahalanobis distances in the formulation of robust affine equivariant estimator of $\boldsymbol{\theta}$, through a tricky ranking of the Mahalanobis distances in (18.5) and an iterative procedure in updating an affine equivariant robust estimator of $\boldsymbol{\theta}$.

The robust estimators in the multivariate case, discussed in detail in Jurečková et al. (2013), are not automatically affine equivariant. With due emphasis on the spatial median, some other estimators were considered by a host of researchers, and we refer to Oja (2010) and Serfling (2010) for a detailed account. Their emphasis is on the spatial median and spatial quantile functions defined as follows:
Let $\mathscr{B}_{p-1}(\mathbf{0})$ be the open unit ball. Then the **u**th *spatial quantile* $Q_F(\mathbf{u})$, $\mathbf{u} \in \mathscr{B}_{p-1}(\mathbf{0})$ is defined as the solution $\mathbf{x} = Q_F(\mathbf{u})$ of the equation

$$\mathbf{u} = \mathbb{E}\left\{ \frac{\mathbf{x} - \mathbf{X}}{\|\mathbf{x} - \mathbf{X}\|} \right\}, \quad \mathbf{u} \in \mathscr{B}_{p-1}(\mathbf{0}).$$

Particularly, $Q_F(\mathbf{0})$ is the spatial median. It is equivariant with respect to $\mathbf{y} = \mathbf{B}\mathbf{x} + \mathbf{b}$, $\mathbf{b} \in \mathbb{R}_p$, $\mathbf{B}$ positive definite and *orthogonal*. However, the spatial quantile function may not be affine equivariant for all $\mathbf{u}$.

Among various approaches to multivariate quantiles we refer to Chakraborty (2001), Roelant and van Aelst (2007), Hallin et al. (2010), Kong and Mizera (2012), Jurečková et al. (2013), among others. Lopuhaä and Rousseeuw (1991) and Zuo (2003, 2004, 2006), among others, studied robust affine equivariant estimators with high breakdown point, based on projection depths. An alternative approach based on the notion of the depth function and associated U-statistics has been initiated by Liu et al. (1999). Notice that every affine invariant function of $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ depends on the $\mathbf{X}_i$ only through the maximal invariant; particularly, this applies to the ranks of the d_{ni} and also to all affine invariant depths considered in the literature. In our formulation, affine-equivariance property is highlighted and accomplished through a ranking of the Mahalanobis distances at various steps.

18.2 Affine Equivariant Linear Estimators

Let $\mathbf{X} \in \mathbb{R}_p$ be a random vector with a distribution function F. Unless stated otherwise, we assume that F is absolutely continuous. Consider the group $\mathscr{C}$ of affine transformations $\mathbf{X} \mapsto \mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{b}$ with $\mathbf{B}$ nonsingular of order $p \times p$, $\mathbf{b} \in \mathbb{R}_p$. Each transformation generates a distribution function G also defined on $\mathbb{R}_p$, which

we denote $G = F_{B,b}$. A vector-valued functional $\boldsymbol{\theta}(F)$, designated as a suitable measure of location of F, is said to be an *affine equivariant location functional*, provided

$$\boldsymbol{\theta}(F_{B,b}) = \mathbf{B}\boldsymbol{\theta}(F) + \mathbf{b} \quad \forall \mathbf{b} \in \mathbb{R}_p,\ \mathbf{B} \text{ positive definite.}$$

Let $\boldsymbol{\Gamma}(F)$ be a matrix valued functional of F, designated as a measure of the *scatter* of F around its location $\boldsymbol{\theta}$ and capturing its *shape* in terms of variation and covariation of the coordinate variables. $\boldsymbol{\Gamma}(F)$ is often termed a *covariance functional*, and a natural requirement is that it is independent of $\boldsymbol{\theta}(F)$. It is termed an *affine equivariant covariance functional*, provided

$$\boldsymbol{\Gamma}(F_{B,b}) = \mathbf{B}\boldsymbol{\Gamma}(F)\mathbf{B}^\top \quad \forall \mathbf{b} \in \mathbb{R}_p,\ \mathbf{B} \text{ positive definite.}$$

We shall construct a class of affine equivariant L-estimators of location parameter, starting with initial affine equivariant location estimator and scale functional, and then iterating them to a higher robustness. For simplicity we start with the sample mean vector $\overline{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$ and with the matrix $\mathbf{V}_n = \mathbf{A}_n^{(0)} = \sum_{i=1}^n (\mathbf{X}_i - \overline{\mathbf{X}}_n)(\mathbf{X}_i - \overline{\mathbf{X}}_n)^\top = n\hat{\boldsymbol{\Sigma}}_n,\ n > p$, where $\hat{\boldsymbol{\Sigma}}_n$ is the sample covariance matrix. Let $R_{ni} = \sum_{j=1}^n I[d_{nj} \le d_{ni}]$ be the rank of d_{ni} among $d_{n1}, \ldots, d_{nn},\ i = 1, \ldots, n$, and denote $\mathbf{R}_n = (R_{n1}, \ldots, R_{nn})^\top$ the vector of ranks. Because F is continuous, the probability of ties is 0, hence the ranks are well defined. Note that d_{ni} and R_{ni} are affine invariant, $i = 1, \ldots, n$. Moreover, the R_{ni} are invariant under any strictly monotone transformation of $d_{ni},\ i = 1, \ldots, n$. Furthermore, each $\mathbf{X}_i$ is trivially affine equivariant. We introduce the following (Mahalanobis) ordering of $\mathbf{X}_1, \ldots, \mathbf{X}_n$:

$$\mathbf{X}_i \prec \mathbf{X}_j \ \Leftrightarrow\ d_{ni} < d_{nj},\ i \ne j = 1, \ldots, n. \tag{18.7}$$

This affine invariant ordering leads to vector of order statistics $\mathbf{X}_{n:1} \prec \ldots \prec \mathbf{X}_{n:n}$ of the sample $\mathbb{X}_n$. In the univariate case with the order statistics $X_{n:1} \le \ldots \le X_{n:n}$, we can consider the k-order *rank-weighted mean* (Sen 1964) defined as

$$T_{nk} = \binom{n}{2k+1}^{-1} \sum_{i=k+1}^{n-k} \binom{i-1}{k}\binom{n-1}{k} X_{n:i}, \quad k = 0, \ldots, \left[\frac{n+1}{2}\right].$$

For $k = 0$, $T_{nk} = \overline{X}_n$ and $k = [(n+1)/2]$ leads to the median $\tilde{X}_n$. In the multivariate case, the ordering is induced by the non-negative d_{ni}, and the smallest d_{ni} corresponds to the smallest *outlyingness* from the center, or to *the nearest*

neighborhood of the center. Keeping that in mind, we can conceive by a sequence $\{k_n\}$ of nonnegative integers, such that k_n is $\nearrow$ in n, but $n^{-1/2}k_n$ is $\searrow$ in n, and for fixed k put

$$\mathbf{L}_{nk} = \binom{k_n}{k}^{-1} \sum_{i=1}^{n} I\left[R_{ni} \le k_n\right] \binom{k_n - R_{ni}}{k-1} \mathbf{X}_i.$$

$\mathbf{L}_{nk}$ is affine equivariant, because the d_{ni} are affine invariant and the $\mathbf{X}_i$ are trivially affine equivariant. Therefore we conclude that the class of L-estimators $\mathbf{L}_{nk}, 0 \le k \le [(n+1)/2]$, are affine equivariant estimators of multivariate location. *This is a characterization of affine-equivariance of the proposed L-estimators*. Of particular interest is the case of $k = 1$, i.e.,

$$\mathbf{L}_{n1} = k_n^{-1} \sum_{i=1}^{n} I\left[R_{ni} \le k_n\right] \mathbf{X}_i$$

representing a trimmed, rank-weighted, nearest neighbor (NN) affine equivariant estimator of $\boldsymbol{\theta}$. In the case $k = 2$ we have

$$\mathbf{L}_{n2} = \binom{k_n}{2}^{-1} \sum_{i=1}^{n} I\left[R_{ni} \le k_n\right] (k_n - R_{ni}) \mathbf{X}_i$$

which can be rewritten as $\mathbf{L}_{n2} = \sum_{i=1}^{n} w_{nR_{ni}} \mathbf{X}_i$ with the weight-function

$$w_{ni} = \begin{cases} \binom{k_n}{2}^{-1} (k_n - i) & \dots\ i = 1, \dots, k_n; \\ 0 & \dots\ i > k_n. \end{cases}$$

We see that $\mathbf{L}_n$ puts greater influence for $R_{ni} = 1$ or 2, and $w_{nk_n} = 0$; $w_{n1} = 2/k_n$. For $k \ge 3$, even greater weights will be given to $R_{ni} = 1$ or 2, etc. For large n we can use the Poisson weights, following Chaubey and Sen (1996):

$$w_{ni}^0 = \left(1 - \mathrm{e}^{-\lambda}\right)^{-1} \frac{\mathrm{e}^{-\lambda} \lambda^i}{i!},\ \lambda < 1,\ i = 1, 2, \dots.$$

A typical λ is chosen somewhere in the middle of $[0, 1]$. Then $\mathbf{L}_n^0 = \sum_{i=1}^{n} w_{nR_{ni}}^0 \mathbf{X}_i$ represents an untrimmed smooth affine equivariant L-estimator of $\boldsymbol{\theta}$; for $\lambda \to 0$, we get the median affine equivariant estimator, while $\lambda \to 1$ gives a version of L_{n2}-type estimator. If λ is chosen close to 1/2 and $k_n = o(\sqrt{n})$, then tail $\sum_{j>k_n} w_{nj}^{(0)}$ converges exponentially to 0, implying a fast negligibility of the tail. Parallelly, the weights $w_n(i)$ can be chosen as the nonincreasing rank scores $a_n(1) \ge a_n(2) \ge \dots \ge a_n(n)$.

To diminish the influence of the initial estimators, we can recursively continue in the same way: Put $\mathbf{L}_n^{(1)} = \mathbf{L}_n$ and define in the next step:

$$\mathbf{A}_n^{(1)} = \sum_{i=1}^{n} (\mathbf{X}_i - \mathbf{L}_n^{(1)})(\mathbf{X}_i - \mathbf{L}_n^{(1)})^\top$$

$$d_{ni}^{(1)} = (\mathbf{X}_i - \mathbf{L}_n^{(1)})^\top (\mathbf{A}_n^{(1)})^{-1} (\mathbf{X}_i - \mathbf{L}_n^{(1)})$$

$$R_{ni}^{(1)} = \sum_{j=1}^{n} I[d_{nj}^{(1)} \le d_{ni}^{(1)}],\ i = 1, \ldots, n, \quad \mathbf{R}_n^{(1)} = (R_{n1}^{(1)}, \ldots, R_{nn}^{(1)})^\top.$$

The second-step estimator is $\mathbf{L}_n^{(2)} = \sum_{i=1}^{n} w_n(R_{ni}^{(1)})\mathbf{X}_i$. In this way we proceed, so at the rth step we define $\mathbf{A}_n^{(r)},\ d_{ni}^{(r)},\ 1 \le i \le n$ and the ranks $\mathbf{R}_n^{(r)}$ analogously, and get the r-step estimator

$$\mathbf{L}_n^{(r)} = \sum_{i=1}^{n} w_n(R_{ni}^{(r-1)})\mathbf{X}_i,\ r \ge 1. \tag{18.8}$$

Note that the $d_{ni}^{(r)}$ are affine invariant for every $1 \le i \le n$ and for every $r \ge 0$. As such, applying an affine transformation $\mathbf{Y}_i = \mathbf{B}\mathbf{X}_i + \mathbf{b},\ \mathbf{b} \in \mathbb{R}_p,\ \mathbf{B}$ positive definite, we see that

$$\mathbf{L}_n^{(r)}(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) = \mathbf{B}\mathbf{L}_n^{(r)}(\mathbf{X}_1, \ldots, \mathbf{X}_n) + \mathbf{b}. \tag{18.9}$$

Hence, the estimating procedure preserves the affine-equivariance at each step and $\mathbf{L}_n^{(r)}$ is an affine equivariant L-estimator of $\boldsymbol{\theta}$ for every r.

The algorithm proceeds as follows:

(1) Calculate $\overline{\mathbf{X}}_n$ and $\mathbf{A}_n^{(0)} = \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}}_n)(\mathbf{X}_i - \overline{\mathbf{X}}_n)^\top$.
(2) Calculate $d_{ni}^{(0)} = (\mathbf{X}_i - \overline{\mathbf{X}}_n)^\top (\mathbf{A}_n^{(0)})^{-1} (\mathbf{X}_i - \overline{\mathbf{X}}_n),\ 1 \le i \le n$.
(3) Determine the rank $R_{ni}^{(0)}$ of $d_{ni}^{(0)}$ among $d_{n1}^{(0)}, \ldots, d_{nn}^{(0)},\ i = 1, \ldots, n$.
(4) Calculate the scores $a_n(i),\ i = 1, \ldots, n$
(5) Calculate the first-step estimator $\mathbf{L}_n^{(1)} = \sum_{i=1}^{n} a_n(R_{ni}^{(0)})\mathbf{X}_i$.
(6) $\mathbf{A}_n^{(1)} = \sum_{i=1}^{n} (\mathbf{X}_i - \mathbf{L}_n^{(1)})(\mathbf{X}_i - \mathbf{L}_n^{(1)})^\top$.
(7) $d_{ni}^{(1)} = (\mathbf{X}_i - \mathbf{L}_n^{(1)})^\top (\mathbf{A}_n^{(1)})^{-1} (\mathbf{X}_i - \mathbf{L}_n^{(1)}),\ 1 \le i \le n$.
(8) $R_{ni}^{(1)}$ = the rank of $d_{ni}^{(1)}$ among $d_{n1}^{(1)}, \ldots, d_{nn}^{(1)},\ i = 1, \ldots, n$.
(9) $\mathbf{L}_n^{(2)} = \sum_{i=1}^{n} a_n(R_{ni}^{(1)})\mathbf{X}_i$.
(10) Repeat the steps (6)–(9).

The estimator $\mathbf{L}_n^{(r)}$ is a linear combination of order statistics corresponding to independent random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$, with random coefficients based on the exchangeable $d_{ni}^{(r)}$. The asymptotic distribution of $\mathbf{L}_n^{(r)}$ under fixed r and for $n \to \infty$ is a problem for a future study, along with the asymptotic distribution of the $d_{ni}^{(r)}$ and of the rank statistics. For the moment, let us briefly recapitulate some asymptotic properties of the $d_{ni}^{(r)}$. Note that $\sum_{i=1}^n d_{ni}^{(r)} = p \quad \forall\ r \geq 0$, and that the $d_{ni}^{(r)}$ are exchangeable nonnegative random variables with a constant sum and $\mathbb{E}(d_{ni}^{(r)}) = \frac{p}{n}$ for every $r \geq 0$. Let $\delta_{ni}^{(r)} = (\mathbf{X}_i - \mathbf{L}_n^{(r)})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \mathbf{L}_n^{(r)})$ and $\delta_i^* = (\mathbf{X}_i - \boldsymbol{\theta})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \boldsymbol{\theta}),\ 1 \leq i \leq n,\ r \geq 1$ where $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{X}_1$. Let $G_n^{(r)}(y) = P\{nd_{ni}^{(r)} \leq y\}$ be the distribution function of the $nd_{ni}^{(r)}$ and let $\hat{G}_n^{(r)}(y) = n^{-1}\sum_{i=1}^n I[nd_{ni}^{(r)} \leq y],\ y \in \mathbb{R}^+$ be the empirical distribution function. Side by side, let $G_{nr}^*(y) = P\{\delta_{ni}^{(r)} \leq y\}$ and $G^*(y) = P\{\delta_i^* \leq y\}$ be the distribution function of $\delta_{ni}^{(r)}$ and δ_i^*, respectively. By the Slutzky theorem,

$$|G_{nr}^*(y) - G^*(y)| \to 0 \ \text{ as } \ n \to \infty.$$

Moreover, by the Courant theorem,

$$\mathrm{Ch}_{min}(\mathbf{AB}^{-1}) = \inf_{\mathbf{x}} \frac{\mathbf{x}^\top \mathbf{Ax}}{\mathbf{x}^\top \mathbf{Bx}} \leq \sup_{\mathbf{x}} \frac{\mathbf{x}^\top \mathbf{Ax}}{\mathbf{x}^\top \mathbf{Bx}} = \mathrm{Ch}_{max}(\mathbf{AB}^{-1}),$$

we have

$$\max_{1\leq i\leq n} \left| \frac{nd_{ni}^{(r)}}{\delta_{ni}^{(r)}} - 1 \right| \leq \max \left\{ \left| \mathrm{Ch}_{max}\left(\frac{1}{n}(\mathbf{A}_n^{(r)})^{-1}\boldsymbol{\Sigma} \right) - 1 \right|, \left| \mathrm{Ch}_{min}\left(\frac{1}{n}(\mathbf{A}_n^{(r)})^{-1}\boldsymbol{\Sigma} \right) - 1 \right| \right\}$$

so that $\left[\frac{1}{n}\mathbf{A}_n^{(r)} \xrightarrow{p} \boldsymbol{\Sigma} \right] \Longrightarrow |\hat{G}_n^{(r)} - G_{nr}^*| \to 0$. In a similar way, $|\delta_{ni}^{(r)} - \delta_i^*| \ll \|\mathbf{L}_{nk}^{(r)} - \boldsymbol{\theta}\|$, where the right-hand side is $O_p(n^{-1/2})$. Because $d_{ni}^{(r)}$ are exchangeable, bounded, and nonnegative random variables, one can use the Hoeffding (1963) inequality to verify that for every $c_n > 0$, there exist positive constants K_1 and ν for which

$$P\left\{ |\hat{G}_n^{(r)}(y) - G_n^{(r)}(y)| > c_n \right\} \leq K_1 \mathrm{e}^{-\nu n c_n^2}.$$

Thus, using $c_n = O\left(n^{1/2}\log n\right)$ can make the right-hand side to converge at any power rate with $n \to \infty$. This leads to the following lemma.

Lemma 18.1 *As* $n \to \infty$,

$$\sup_{d\in\mathbb{R}^+} \left\{ |\hat{G}_n^{(r)}(y) - G_n^{(r)}(y) - \hat{G}_n^{(r)}(y') + G_n^{(r)}(y')| : \ |y - y'| \le n^{-1/2}\sqrt{2\log n} \right\}$$
$$\stackrel{a.s.}{=} O(n^{-\frac{3}{4}} \log n). \tag{18.10}$$

Proof (Outline) The lemma follows from the Borel–Cantelli lemma, when we notice that both $\hat{G}_n^{(r)}(y)$ and $G_n^{(r)}(y)$ are $\nearrow$ in $y \in \mathbb{R}^+$, and that $\hat{G}_n^{(r)}(0) = G_n^{(r)}(0)$, $\hat{G}_n^{(r)}(\infty) = G_n^{(r)}(\infty) = 1$.

Theorem 18.1 *Let*

$$W_{nr}(t) = n^{-1/2}[\hat{G}_n^{(r)}(G_n^{(r)-1}(t)) - t], \ t \in [0,1]; \ W_{nr} = \{W_{nr}(t); \ 0 \le t \le 1\}.$$

Then $W_{nr} \Rightarrow W$ *in the Skorokhod* $\mathscr{D}[0,1]$ *topology, where* W *is a Brownian Bridge on [0,1].*

Proof (outline) The tightness part of the proof follows from Lemma 18.1. For the convergence of finite-dimensional distributions, we appeal to the central limit theorem for interchangeable random variables of Chernoff and Teicher (1958).

If the $\mathbf{X}_i$ have multinormal distribution, then δ_i^* has the Chi-squared distribution with p degrees of freedom. If $\mathbf{X}_i$ is elliptically symmetric, then its density depends on $h(\|\mathbf{x} - \boldsymbol{\theta}\|_\Sigma)$, with $h(y), \ y > 0$ depending only on the norm $\|y\|$. If $p \ge 2$, as it is in our case, it may be reasonable to assume that $H(y) = \int_0^y h(u)du$ behaves as $\sim y^{p/2}$ (or higher power) for $y \to 0$. Thus $y \sim [H(y)]^{2/p}$ (or $[H(y)]^r, \ r \le 2/p$) for $y \to 0$. On the other hand, since our choice is $k_n = o(n)$, the proposed estimators $\mathbf{L}_{n1}$ and $\mathbf{L}_{n2}$ both depend on the $\mathbf{X}_i$ with d_{ni} of lower rank ($R_{ni} \le k_n$ or $n^{-1}R_{ni} \le n^{-1}k_n \to 0$). Hence, both $\mathbf{L}_{n1}$ and $\mathbf{L}_{n2}$ are close to the induced vector $\mathbf{X}_{[1]}$ where $[1] = \{i : \ R_{ni} = 1\}$. If the initial estimator is chosen as $\mathbf{X}_{[1]}$ and $\mathbf{A}_{n[1]} = n^{-1}\sum_{i=1}^n (\mathbf{X}_i - \overline{\mathbf{X}}_{[1]})(\mathbf{X}_i - \overline{\mathbf{X}}_{[1]})^\top$, then the iteration process will be comparatively faster than if we start with the initial estimators $\overline{\mathbf{X}}_n$ and $n^{-1}\mathbf{A}_n^{(0)}$.

The proposed $\mathbf{L}_{n1}, \ \mathbf{L}_{n2}$ are both affine equivariant and robust. If we define the D-efficiency

$$\mathbf{D}_n^{(r)} = \left(\frac{|\mathbf{A}_n^{(r)}|}{|\mathbf{A}_n^{(0)}|} \right)^{1/p}, \quad r \ge 1, \tag{18.11}$$

then it will be slightly better than that of the spatial median; the classical $\overline{\mathbf{X}}_n$ has the best efficiency for multinormal distribution but it is much less robust than $\mathbf{L}_{n1}$ and $\mathbf{L}_{n2}$.

18.3 Numerical Illustration

18.3.1 Multivariate Normal Distribution

The procedure is illustrated on samples of size $n = 100$ simulated from the multivariate normal distribution $\mathcal{N}_3(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{bmatrix} \tag{18.12}$$

and each time the affine equivariant trimmed $\mathbf{L}_{n1}$-estimator ($k_n = 15$) and affine equivariant $\mathbf{L}_{n2}$-estimator were calculated in 10 iterations of the initial estimator. 5000 replications of the model were simulated and also the mean was computed, for the sake of comparison. Results are summarized in Table 18.1. Figure 18.1 illustrates the distribution of estimated parameters θ_1, θ_2, θ_3 for various iterations of $\mathbf{L}_{n1}$-estimator and $\mathbf{L}_{n2}$-estimator and compares them with the mean and median. Tables 18.2 and 18.3 and Fig. 18.2 compare the D-efficiency of proposed estimators.

The Mahalanobis distance is also illustrated. One sample of size $n = 100$ was simulated from the bivariate normal distribution with the above parameters. Afterwards, the Mahalanobis distances $d_{ii} = (\mathbf{X}_i - \overline{\mathbf{X}}_n)^T \mathbf{S}_n^{-1} (\mathbf{X}_i - \overline{\mathbf{X}}_n)$, $i = 1, \ldots, n$ were calculated. They represent n co-axial ellipses centered at $\overline{\mathbf{X}}_n$—see

Table 18.1 Normal distribution: the mean in the sample of 5000 replications of estimators $\mathbf{L}_{n1}$ (trimmed) and $\mathbf{L}_{n2}$, sample sizes $n = 100$

i	$\mathbf{L}_{n1}^{(i)}$			$\mathbf{L}_{n2}^{(i)}$		
1	0.999607	2.001435	−0.998297	0.999401	1.999796	−0.999943
2	0.999473	2.001423	−0.996185	0.999083	1.999584	−0.999782
3	0.999519	2.000290	−0.993274	0.998926	1.999509	−0.999801
4	0.999435	2.000190	−0.991901	0.998854	1.999496	−0.999871
5	0.999474	2.000771	−0.991295	0.998811	1.999476	−0.999973
6	0.999646	2.001285	−0.990964	0.998781	1.999471	−1.000032
7	0.999926	2.001519	−0.990829	0.998773	1.999470	−1.000049
8	0.999952	2.001529	−0.990803	0.998772	1.999472	−1.000068
9	0.999939	2.001497	−0.990745	0.998775	1.999489	−1.000071
10	0.999853	2.001424	−0.990711	0.998779	1.999497	−1.000061

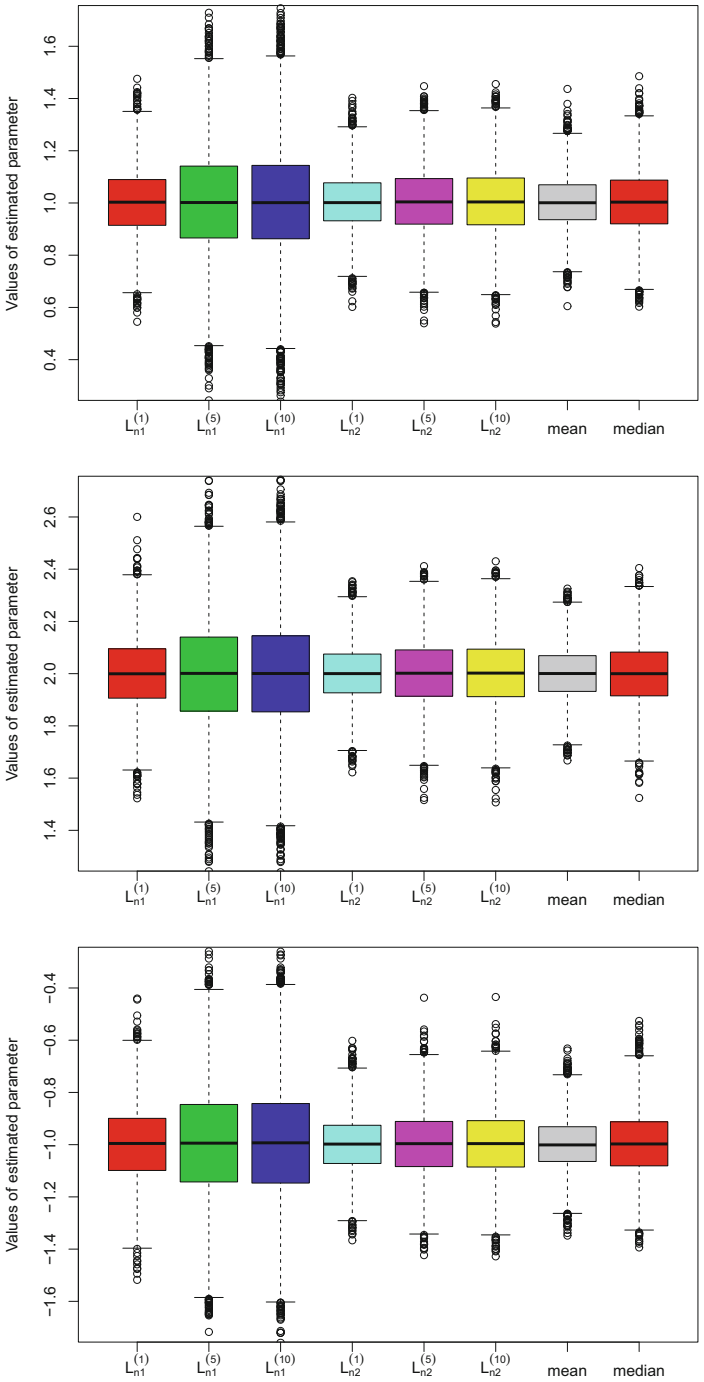

Fig. 18.1 Normal distribution: box-plots of the 5000 estimated values of $\theta_1(=1)$ (*top*), $\theta_2(=2)$ (*middle*), and $\theta_3(=-1)$ (*bottom*) for the $\mathbf{L}_{n1}^{(1)}$, $\mathbf{L}_{n1}^{(5)}$, $\mathbf{L}_{n1}^{(10)}$, $\mathbf{L}_{n2}^{(1)}$, $\mathbf{L}_{n2}^{(5)}$, $\mathbf{L}_{n2}^{(10)}$, mean and median

Table 18.2 Normal distribution: the mean, median, and minimum of D-efficiency in the sample of 5000 replications of estimators $\mathbf{L}_{n1}$ (trimmed) and $\mathbf{L}_{n2}$, sample sizes $n = 100$

	$\mathbf{L}_{n1}^{(i)}$			$\mathbf{L}_{n2}^{(i)}$		
Iteration	Mean	Median	Minimum	Mean	Median	Minimum
2	1.000637	0.999615	0.871029	0.999862	0.999670	0.951057
3	1.001272	0.998593	0.818591	0.999625	0.999451	0.946809
4	1.001106	0.997822	0.793030	0.999392	0.999231	0.946575
5	1.000697	0.997283	0.794157	0.999205	0.999041	0.946332
6	1.000394	0.997211	0.793903	0.999078	0.998886	0.946022
7	1.000131	0.997122	0.793903	0.998997	0.998802	0.945613
8	0.999916	0.996964	0.793903	0.998951	0.998807	0.945032
9	0.999882	0.996924	0.793903	0.998921	0.998768	0.944518
10	0.999860	0.996924	0.793903	0.998899	0.998706	0.944272

Table 18.3 Normal distribution: the 25 %-quantile, 75 %-quantile and max of D-efficiency in the sample of 5000 replications of estimators $\mathbf{L}_{n1}$ (trimmed) and $\mathbf{L}_{n2}$, sample sizes $n = 100$

	$\mathbf{L}_{n1}^{(i)}$			$\mathbf{L}_{n2}^{(i)}$		
Iteration	25%-quantile	75%-quantile	Max	25%-quantile	75%-quantile	Max
2	0.971669	1.028444	1.169137	0.990914	1.008819	1.057115
3	0.960891	1.039437	1.295245	0.989675	1.009489	1.065753
4	0.956068	1.042774	1.294952	0.989358	1.009426	1.068542
5	0.953186	1.043542	1.301147	0.989098	1.009239	1.069236
6	0.952193	1.044435	1.305327	0.989000	1.009182	1.069437
7	0.951535	1.044942	1.326996	0.988916	1.009142	1.069600
8	0.951441	1.044394	1.330791	0.988839	1.009109	1.069226
9	0.951452	1.044562	1.330791	0.988802	1.009087	1.069236
10	0.951356	1.044749	1.330791	0.988771	1.009062	1.069278

Fig. 18.3 (black ellipses). The modified Mahalanobis distances replaced $\overline{\mathbf{X}}_n$ by the affine equivariant trimmed $\mathbf{L}_{n1}$-estimator with $k_n = 15$ (see the blue ellipses on Fig. 18.3) and affine equivariant $\mathbf{L}_{n2}$-estimator (see the red ellipses on Fig. 18.3) with analogous modification of $\mathbf{S}_n$.

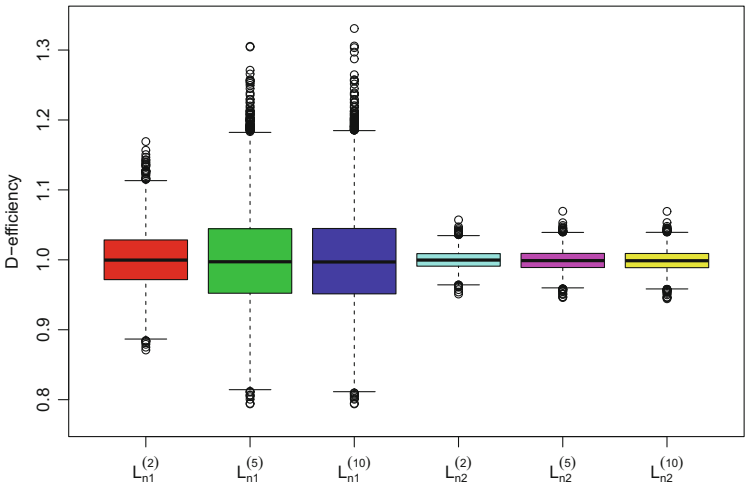

Fig. 18.2 Normal distribution: box-plots of the 5000 estimated values of D-efficiency for the $\mathbf{L}_{n1}^{(2)}$, $\mathbf{L}_{n1}^{(5)}$, $\mathbf{L}_{n1}^{(10)}$, $\mathbf{L}_{n2}^{(2)}$, $\mathbf{L}_{n2}^{(5)}$, $\mathbf{L}_{n2}^{(10)}$

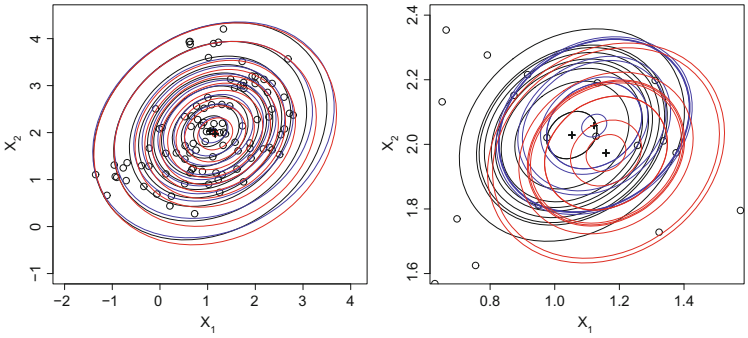

Fig. 18.3 Normal distribution: Mahalanobis distances represented by co-axial ellipses centered at the mean $\bar{X}$ (*black*), at the trimmed $\mathbf{L}_{n1}$-estimator (*blue*) and at the $\mathbf{L}_{n2}$-estimator (*red*). All simulated bivariate data with the every tenth contour are illustrated on the *left*, detail of the center with the first ten contours is on the *right*

Table 18.4 t-distribution: the mean in the sample of 5000 replications of estimators $\mathbf{L}_{n1}$ (trimmed) and $\mathbf{L}_{n2}$, sample sizes $n = 100$

i	$\mathbf{L}_{n1}^{(i)}$			$\mathbf{L}_{n2}^{(i)}$		
1	1.004760	2.002252	−0.992618	1.003081	2.001199	−0.998781
2	1.005034	2.001851	−0.991150	1.002154	2.000044	−0.999996
3	1.005473	2.001430	−0.991330	1.001744	1.999661	−1.000627
4	1.005334	2.000732	−0.992535	1.001594	1.999526	−1.000909
5	1.005351	2.000951	−0.993532	1.001540	1.999475	−1.001028
6	1.005007	2.001127	−0.994361	1.001528	1.999452	−1.001069
7	1.004636	2.001009	−0.994817	1.001522	1.999447	−1.001086
8	1.004520	2.000930	−0.995011	1.001524	1.999443	−1.001090
9	1.004466	2.000911	−0.995172	1.001526	1.999442	−1.001090
10	1.004445	2.000821	−0.995221	1.001527	1.999444	−1.001086

18.3.2 Multivariate t-Distribution

Similarly, we illustrate the procedure on samples of size $n = 100$ simulated from the multivariate t-distribution with 3 degrees of freedom $t_3(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, with the same parameters as in (18.12). Each time, ten iterations of affine equivariant trimmed $\mathbf{L}_{n1}$-estimator ($k_n = 15$) and of affine equivariant $\mathbf{L}_{n2}$-estimator, started from the initial estimator, were calculated. Five thousand replications of the model were simulated and the mean was computed, for the sake of comparison. Results are summarized in Table 18.4. Figure 18.4 illustrates the distribution of estimated parameters θ_1, θ_2, θ_3 for various iterations of $\mathbf{L}_{n1}$-estimator and $\mathbf{L}_{n2}$-estimator and compares them with the mean and median. Tables 18.5 and 18.6 and Fig. 18.5 compare the D-efficiencies of the proposed estimators and Fig. 18.6 illustrates the Mahalanobis distances.

Although $\mathbf{L}_n^{(1)}$ resembles the NN-estimator, its behavior for t-distribution reveals its robustness no less than $\mathbf{L}_n^{(2)}$. For multivariate normal distribution, both $\mathbf{L}_n^{(1)}$ and $\mathbf{L}_n^{(2)}$ seem to be doing well against outliers. Figures 18.3 and 18.6 illustrate this feature in a visible way.

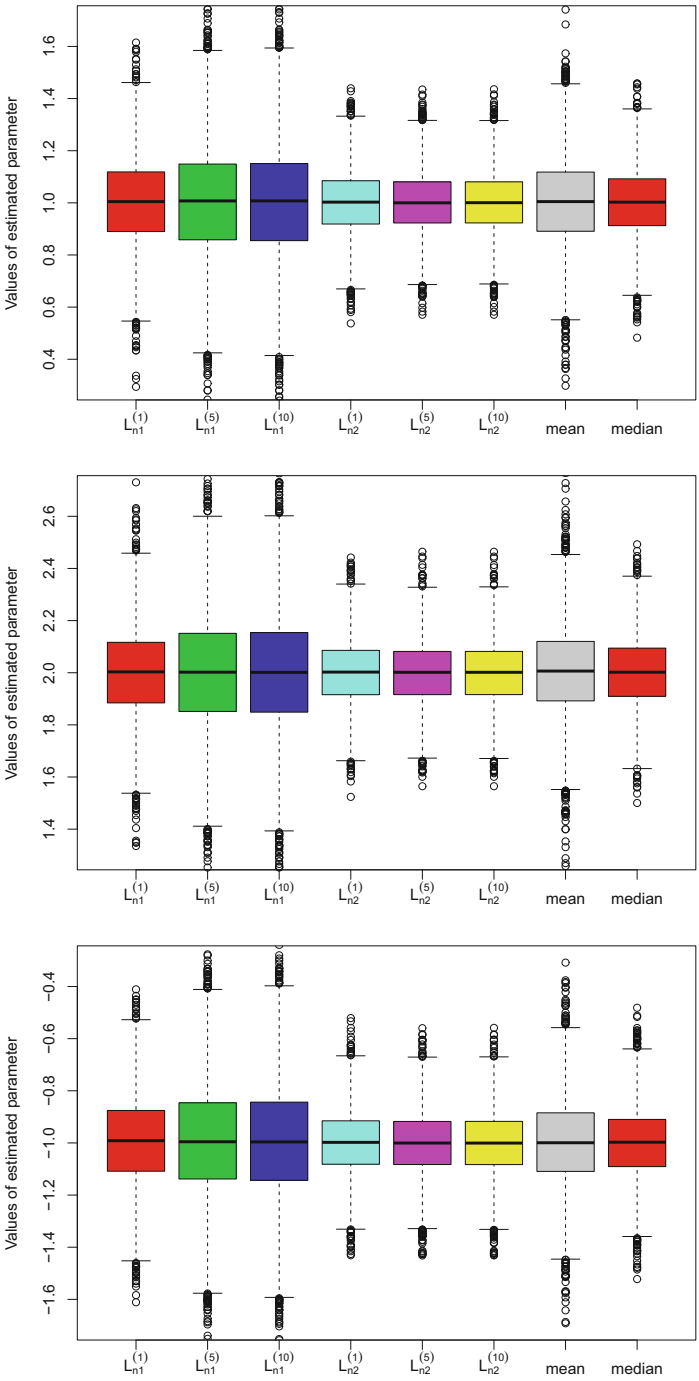

Fig. 18.4 *t*-distribution: box-plots of the 5000 estimated values of $\theta_1(=1)$ (*top*), $\theta_2(=2)$ (*middle*), and $\theta_3(=-1)$ (*bottom*) for the $\mathbf{L}_{n1}^{(1)}$, $\mathbf{L}_{n1}^{(5)}$, $\mathbf{L}_{n1}^{(10)}$, $\mathbf{L}_{n2}^{(1)}$, $\mathbf{L}_{n2}^{(5)}$, $\mathbf{L}_{n2}^{(10)}$, mean and median

Table 18.5 t-distribution: the mean, median, and minimum of D-efficiency in the sample of 5000 replications of estimators $\mathbf{L}_{n1}$ (trimmed) and $\mathbf{L}_{n2}$, sample sizes $n = 100$

	$\mathbf{L}_{n1}^{(i)}$			$\mathbf{L}_{n2}^{(i)}$		
Iteration	Mean	Median	Minimum	Mean	Median	Minimum
2	1.001813	1.000857	0.896048	1.000812	1.000303	0.857210
3	1.001827	1.001157	0.824105	1.000556	1.000109	0.845643
4	1.001377	1.000619	0.810082	1.000360	0.999912	0.844578
5	1.000887	0.999840	0.796776	1.000260	0.999766	0.844182
6	1.000372	0.999121	0.777458	1.000214	0.999723	0.843850
7	1.000024	0.999086	0.756777	1.000196	0.999740	0.843933
8	0.999881	0.998921	0.756555	1.000187	0.999714	0.843928
9	0.999806	0.998907	0.756655	1.000184	0.999726	0.843917
10	0.999753	0.998921	0.756655	1.000183	0.999726	0.843897

Table 18.6 t-distribution: the 25 %-quantile, 75 %-quantile and max of D-efficiency in the sample of 5000 replications of estimators $\mathbf{L}_{n1}$ (trimmed) and $\mathbf{L}_{n2}$, sample sizes $n = 100$

	$\mathbf{L}_{n1}^{(i)}$			$\mathbf{L}_{n2}^{(i)}$		
Iteration	25 %-quantile	75 %-quantile	Max	25 %-quantile	75 %-quantile	Max
2	0.980004	1.023251	1.157852	0.986053	1.014785	1.194658
3	0.972157	1.030707	1.212343	0.983882	1.016369	1.212103
4	0.967982	1.032541	1.214165	0.983518	1.016634	1.214597
5	0.967034	1.032940	1.222945	0.983380	1.016629	1.215499
6	0.966116	1.032850	1.260563	0.983366	1.016565	1.215862
7	0.965784	1.032895	1.261640	0.983391	1.016572	1.216013
8	0.965561	1.032746	1.261675	0.983406	1.016537	1.216332
9	0.965436	1.032643	1.261675	0.983380	1.016522	1.216431
10	0.965404	1.032680	1.261675	0.983406	1.016522	1.216579

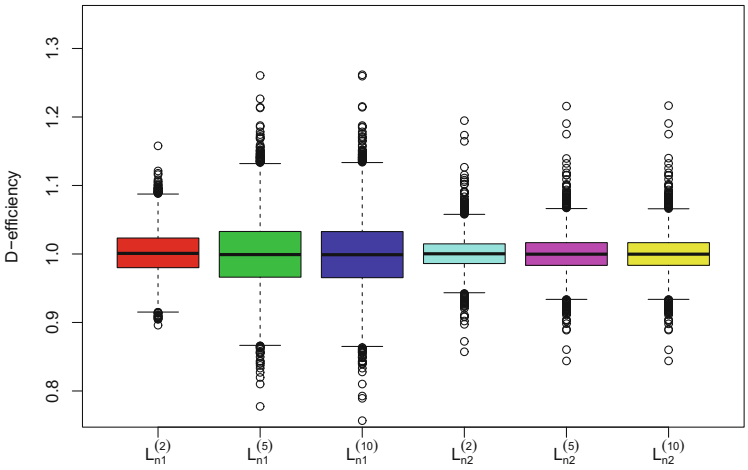

Fig. 18.5 *t*-distribution: box-plots of the 5000 estimated values of D-efficiency for the $\mathbf{L}_{n1}^{(2)}$, $\mathbf{L}_{n1}^{(5)}$, $\mathbf{L}_{n1}^{(10)}$, $\mathbf{L}_{n2}^{(2)}$, $\mathbf{L}_{n2}^{(5)}$, $\mathbf{L}_{n2}^{(10)}$

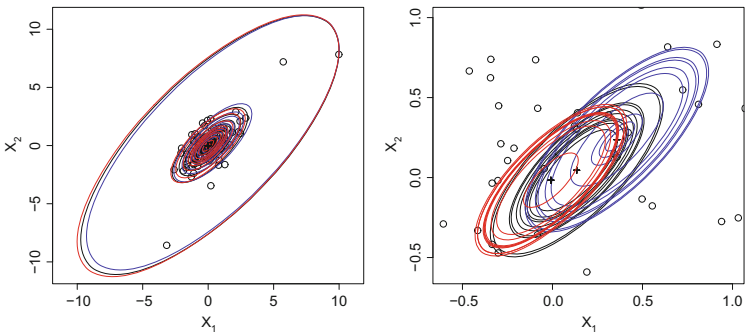

Fig. 18.6 *t*-distribution: Mahalanobis distances represented by co-axial ellipses centered at the mean $\bar{X}$ (*black*), at the trimmed $\mathbf{L}_{n1}$-estimator (*blue*) and at the $\mathbf{L}_{n2}$-estimator (*red*). All simulated bivariate data with the every tenth contour are illustrated on the *left*, detail of the center with the first ten contours is on the *right*

Acknowledgements We highly appreciate the long-term cooperation with Hannu Oja and his group, and hope in new fresh interesting joint ideas in the future. We also thank the Editors for the organization of this Volume.

The authors thank two Referees for their comments, which helped to better understanding the text. P. K. Sen gratefully acknowledges the support of the Cary C. Boshamer Research Foundation at UNC. Research of J. Jurečková and of J. Picek was supported by the Czech Republic Grant 15–00243S.

References

Chakraborty, B.: On affine equivariant multivariate quantiles. Ann. Inst. Stat. Math. **53**, 380–403 (2001)

Chaubey, Y.P., Sen, P.K.: On smooth estimation of survival and density functions. Stat. Decis. **14**, 1–22 (1996)

Chernoff, H., Teicher, H.: A central limit theorem for sums of interchangeable random variables. Ann. Math. Stat. **29**, 118–130 (1958)

Eaton, M.: Multivariate Statistics. Wiley, New York (1983)

Hallin, M., Paindaveine, D., Šíman, M.: Multivariate quantiles and multiple-output regression quantiles: From L1 optimization to halfspace depth. Ann. Stat. **38**, 635–669 (2010)

Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**, 13–30 (1963)

Jurečková, J., Sen, P.K., Picek, J.: Methodology in Robust and Nonparametric Statistics. Chapman & Hall/CRC, Boca Raton (2013)

Kong, L., Mizera, I.: Quantile tomography: using quantiles with multivariate data. Stat. Sin. **22**, 1589–1610 (2012)

Liu, R.Y., Parelius, J.M., Singh, K.: Multivariate analysis by data depth: descriptive statistics, graphics and inference. Ann. Stat. **27**, 783–858 (1999)

Lopuhaä, H.P., Rousseeuw, P.J.: Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. Ann. Stat. **19**, 229–248 (1991)

Obenchain, R.L.: Multivariate procedures invariant under linear transformations. Ann. Math. Stat. **42**, 1569–1578 (1971)

Oja, H.: Multivariate Nonparametric Methods with R. An Approach Based on Spatial Signs and Ranks. Lecture Notes in Statistics, vol. 199, Springer, New York (2010)

Roelant, E., Van Aelst, S.: An L1-type estimator of multivariate location and shape. Stat. Methods Appl. **15**, 381–393 (2007)

Sen, P.K.: On some properties of rank-weighted means. J. Indian Soc. Agric. Stat. **16**, 51–61 (1964)

Serfling, R.J.: Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardization. J. Nonparametric Stat. **22**, 915–936 (2010)

Weber, N.C.: A martingale approach to central limit theorems for exchangeable random variables. J. Appl. Probab. **17**, 662–673 (1980)

Zuo, Y.: Projection-based depth functions and associated medians. Ann. Stat. **31**, 1460–1490 (2003)

Zuo, Y.: Projection-based affine equivariant multivariate location estimators with the best possible finite sample breakdown point. Stat. Sin. **14**, 1199–2008 (2004)

Zuo, Y.: Multidimensional trimming based on projection depth. Ann. Stat. **34**, 2211–2251 (2006)

Chapter 19
Robust High-Dimensional Precision Matrix Estimation

Viktoria Öllerer and Christophe Croux

Abstract The dependency structure of multivariate data can be analyzed using the covariance matrix $\boldsymbol{\Sigma}$. In many fields the precision matrix $\boldsymbol{\Sigma}^{-1}$ is even more informative. As the sample covariance estimator is singular in high dimensions, it cannot be used to obtain a precision matrix estimator. A popular high-dimensional estimator is the graphical lasso, but it lacks robustness. We consider the high-dimensional independent contamination model. Here, even a small percentage of contaminated cells in the data matrix may lead to a high percentage of contaminated rows. Downweighting entire observations, which is done by traditional robust procedures, would then result in a loss of information. In this paper, we formally prove that replacing the sample covariance matrix in the graphical lasso with an elementwise robust covariance matrix leads to an elementwise robust, sparse precision matrix estimator computable in high dimensions. Examples of such elementwise robust covariance estimators are given. The final precision matrix estimator is positive definite, has a high breakdown point under elementwise contamination, and can be computed fast.

Keywords Cellwise contamination • Covariance estimation • Elementwise robustness • Graphical modeling • Independent contamination model • Sparse

19.1 Introduction

Many statistical methods that deal with the dependence structures of multivariate data sets start from an estimate of the covariance matrix. For observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ with $n > p$, the classical sample covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top, \tag{19.1}$$

where $\bar{\mathbf{x}} \in \mathbb{R}^p$ denotes the mean of the data, is optimal in many ways. It is easy to compute, maximizes the likelihood function for normal data, is unbiased and

V. Öllerer • C. Croux (✉)
Faculty of Economics and Business, KU Leuven, 3000 Leuven, Belgium
e-mail: viktoria.oellerer@kuleuven.be; christophe.croux@kuleuven.be

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_19

consistent. However, problems arise when p increases. For $p \approx n$, the sample covariance matrix has low precision and for $p > n$ it even becomes singular, such that the estimated precision matrix $\hat{\boldsymbol{\Theta}} := \hat{\boldsymbol{\Sigma}}^{-1}$ is not computable anymore. This is a problem since there are many fields where the precision matrix is needed rather than the covariance matrix. Computation of Mahalanobis distances or linear discriminant analysis are just two examples. The most popular field using precision matrices is probably Gaussian graphical modeling, where the nodes of the graph represent the different variables. If an element $(\hat{\boldsymbol{\Theta}})_{ij}$ of the estimated precision matrix equals zero, the variables i and j are independent given all the other variables, and no edge is drawn between the nodes representing variables i and j. Therefore, edges correspond to nonzero elements of the precision matrix. As a result, the whole graph can be recovered if the support of the precision matrix is known. This leads to an increasing interest in sparse precision matrices (precision matrices with a lot of zero elements) as interpretation of the graph will be eased if the number of nonzeros in the precision matrix is kept small.

The three most suitable techniques to compute sparse precision matrices that are also applicable in high dimensions are the graphical lasso (GLASSO) (Friedman et al. 2008), the quadratic approximation method for sparse inverse covariance learning (QUIC) (Hsieh et al. 2011) and the constrained L_1-minimization for inverse matrix estimation (CLIME) (Cai et al. 2011). All three methods start from the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ and try to minimize a criterion based on the log-likelihood (see Sect. 19.2). Since these estimators use the nonrobust sample covariance matrix as an input, they are only suitable for clean data that do not contain any outliers.

The problem, however, is that data is rarely clean. Thus, there is need for robust procedures. Most robust procedures downweight observations as a whole ("rowwise downweighting"). However, in many statistical applications only a limited number of observations are available, while large amounts of variables are measured for each observation. Downweighting an entire observation because of one single outlying cell in the data matrix results in a huge loss of information. Additionally, the contaminating mechanism may be independent for different variables. In this case, the probability of having an observation without contamination in any cell is decreasing exponentially when the number of variables increases. As an example, imagine a data set, where each observation contains exactly one contaminated cell. Even though there is not a single fully clean observation, each observation still contains a lot of clean information. Nonetheless, the "classical" robust procedures (that downweight whole observations) cannot deal with a data set like that, since they need at least half of the observations to be absolutely clean of contamination. This type of "cellwise" or "elementwise" contamination was formally described by Alqallaf et al. (2009), who extend the usual Tukey–Huber contamination model (the model that considers whole observations as either outlying or clean). In this more extensive setup, a random vector

$$\mathbf{x} = (\mathbf{I}_p - \mathbf{B})\mathbf{y} + \mathbf{B}\mathbf{z}$$

is observed, where $\mathbf{y}$ follows the model distribution and $\mathbf{z}$ some arbitrary distribution creating contamination, and $\mathbf{y}$, $\mathbf{B}$, and $\mathbf{z}$ are independent. Depending on the Bernoulli random variables B_i with $\mathbb{P}[B_i = 1] = \epsilon_i$ that build the diagonal matrix $\mathbf{B} = \text{diag}(B_1, \ldots, B_p)$, different types of outliers are created. If all B_i are independent $(i = 1, \ldots, p)$, we speak about "cellwise contamination." If $\mathbb{P}[B_1 = B_2 = \ldots = B_p] = 1$, rowwise contamination is created. Under any type of contamination, the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ is not a good estimator anymore, as it can be distorted by just a single outlying observation.

For robust covariance matrix estimation under rowwise contamination, a lot of work has been done. One of the most popular rowwise robust covariance estimators is the minimum covariance determinant (Rousseeuw and Van Driessen 1999). It has a high breakdown point and is very fast to compute. However, it is not computable in high dimensions. Another estimator with very nice theoretical properties is the affine equivariant rank covariance matrix (Ollila et al. 2003). It is very efficient and has maximal breakdown point. However, its computation is extremely time consuming, especially in high-dimensions. Maronna and Zamar (2002) propose a high-dimensional covariance estimator, an orthogonalized version of the Gnanadesikan-Kettenring estimate (OGK). Another very simple estimator has been developed by Visuri et al. (2000). Their spatial sign covariance matrix appeals through a simple definition and can be computed very fast, even in high dimensions. Very recently, Ollila and Tyler (2014) have introduced a regularized M-estimator of scatter. Under general conditions, they show existence and uniqueness of the estimator, using the concept of geodesic convexity.

Much less work has been done for covariance estimation under cellwise contamination. A first approach was taken by Van Aelst et al. (2011), who defined a cellwise weighting scheme for the Stahel–Donoho estimator. However, as for the original estimate, computation times are not feasible for larger numbers of variables. A very recent approach by Agostinelli et al. (2015) flags cellwise outliers as missing values and applies afterwards a rowwise robust method that can deal with missing values. By this, it can deal with cellwise and rowwise outliers at the same time, but again, computation for high-dimensions is not achievable.

The first step to deal with cellwise outliers in very high dimensions has been taken by Alqallaf et al. (2002). They first compute a pairwise correlation matrix. Afterwards the OGK estimate is applied to obtain a positive semidefinite covariance estimate. This method has been fine-tuned by Tarr et al. (2015) who use pairwise covariances instead of correlations (see also Tarr 2014). This matrix is then plugged into the graphical lasso (and similar techniques) instead of the sample covariance matrix, resulting in a sparse precision matrix estimate. A very different approach has been taken by Finegold and Drton (2011). Replacing the assumption of Gaussian distribution of the data with t-distribution gives more robust results since the t-distribution has heavier tails. Assuming a so-called alternative t-distribution (see Sect. 19.6) results in robustness against cellwise contamination.

In this paper, we consider different high-dimensional precision matrix estimators robust to cellwise contamination. Our approach is similar in spirit as Tarr et al.

(2015) (see also Tarr 2014), but we emphasize the difference in Sect. 19.3. We start with pairwise robust correlation estimates from which we then estimate a covariance matrix by multiplication with robust standard deviations. This cellwise robust covariance matrix replaces then the sample covariance matrix in the GLASSO, yielding a sparse, cellwise robust precision matrix estimator. The different nonrobust precision matrix estimators are introduced in Sect. 19.2. The cellwise robust covariance matrix estimators are explained in Sect. 19.3. We discuss the selection of the regularization parameter in Sect. 19.4. In Sect. 19.5, the breakdown point of the proposed precision matrix estimator is derived. Simulation studies are presented in Sect. 19.6. In Sect. 19.7, we discuss possible applications of the proposed method and present a real data example. Section 19.8 concludes.

19.2 High-Dimensional Sparse Precision Matrix Estimation for Clean Data

Recently, a lot of effort has been put into designing estimators and efficient routines for high-dimensional precision matrix estimation. We focus here on sparse precision matrix estimation, that is, procedures that result in a precision matrix containing many zero elements. In this section, we review three techniques that start from an estimate of the covariance matrix $\hat{\boldsymbol{\Sigma}}$ and then optimize a criterion based on the likelihood function to find the precision matrix estimate. Since the methods are based on the sample covariance matrix, they are only useful if no contamination is present in the data.

The graphical lasso (GLASSO) (Friedman et al. 2008) maximizes the L_1-penalized log-likelihood function:

$$\hat{\boldsymbol{\Theta}}_{GL}(\mathbf{X}) = \underset{\substack{\boldsymbol{\Theta}\in\mathbb{R}^{p\times p}\\ \boldsymbol{\Theta}\succ 0}}{\arg\max} \log\det(\boldsymbol{\Theta}) - \operatorname{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}) - \rho\sum_{j,k=1}^{p}|(\Theta)_{jk}|, \tag{19.2}$$

where $\mathbf{A} \succ 0$ denotes a strictly positive definite matrix $\mathbf{A}$ and ρ is a regularization parameter. If the regularization parameter ρ is equal to zero, the solution of the GLASSO is the inverse of the sample covariance matrix. The larger the value of ρ is chosen, the more sparse the precision matrix estimate becomes. Since the objective function (19.2) is concave, there exists a unique solution. Banerjee et al. (2008) showed that the solution of the GLASSO always results in a strictly positive definite estimate $\hat{\boldsymbol{\Theta}}_{GL}(\mathbf{X})$ for any $\rho > 0$, even if $p > n$, and this for any positive semidefinite, symmetric matrix $\hat{\boldsymbol{\Sigma}}$ in (19.2).

The solution $\hat{\boldsymbol{\Theta}}_{GL}(\mathbf{X})$ can be computed via iterative multiple lasso regression in a block coordinate descent fashion. That means that each column of the final estimate is computed separately. Looking at the first order condition only for the target column, the equation can be seen as a first order condition of a multiple lasso regression. The GLASSO algorithm loops through all columns of the precision matrix

iteratively, computing each time the multiple lasso regression, until convergence of the precision matrix estimate is reached. Note that the algorithm does not use the data directly, but only uses it indirectly by using the sample covariance matrix. The GLASSO algorithm is implemented in Fortran and available through the R-package `glasso` (Friedman et al. 2014). However, this implementation sometimes encounters convergence problems. Therefore, we use in the remainder of this paper the implementation of the GLASSO algorithm in the R-package `huge` (Zhao et al. 2014), where these convergence issues have been solved.

Another algorithm solving (19.2) is the quadratic approximation method for sparse inverse covariance learning (QUIC) (Hsieh et al. 2011). Both the GLASSO algorithm and the QUIC compute a solution for the same objective function. It turned out that QUIC was performing considerably slower in high dimensions than the GLASSO implementation in the R-package huge (Zhao et al. 2014), and therefore we will not deal with the former in this paper.

The constrained L_1-minimization for inverse matrix estimation (CLIME) is defined as

$$\hat{\boldsymbol{\Theta}}_1(\mathbf{X}) = \arg\min_{\boldsymbol{\Theta}\in\mathbb{R}^{p\times p}} \sum_{i=1}^{p}\sum_{j=1}^{p} |\boldsymbol{\Theta}_{ij}| \qquad \text{subject to } \max_{\substack{i=1,\ldots,p\\ j=1,\ldots,p}} |(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Theta} - \mathbf{I}_p)_{ij}| \le \rho$$

$$\hat{\boldsymbol{\Theta}}_C(\mathbf{X}) = (\hat{\theta}_{ij}) \quad \text{with} \quad \hat{\theta}_{ij} = \hat{\theta}^1_{ij} I_{[|\hat{\theta}^1_{ij}|\le|\hat{\theta}^1_{ji}|]} + \hat{\theta}^1_{ji} I_{[|\hat{\theta}^1_{ij}|>|\hat{\theta}^1_{ji}|]} \quad \text{and} \quad \hat{\Theta}_1(\mathbf{X}) = (\hat{\theta}^1_{ij}).$$

The result is a symmetric matrix that is positive definite with high probability. The CLIME estimator $\hat{\boldsymbol{\Theta}}_C(\mathbf{X})$ converges fast towards the true precision matrix under some mild conditions. The algorithm is implemented in the R-package `clime` (Cai et al. 2012). Like the GLASSO algorithm, it does not use the data directly, but only requires the sample covariance matrix as an input. Replacing the sample covariance matrix with a cellwise robust estimator (see Sect. 19.3), the resulting estimator is similarly accurate (with respect to Kullback–Leibler divergence measure, see Sect. 19.6) as the one obtained when plugging the cellwise robust estimator into the GLASSO estimator. In some cases, plugging the robust estimator into the CLIME led to slightly better accuracy. However, the computation time was much higher than when plugging it into the GLASSO (for $p = 60$ the computation time was more than ten times higher). Since in high-dimensional analysis computation time is important, we will not consider this estimator in the remainder of the paper.

19.3 Cellwise Robust, Sparse Precision Matrix Estimators

We start with computing a cellwise robust covariance matrix $\mathbf{S}$ by pairwise, robust estimation of the covariances. This cellwise robust covariance matrix can then be used to replace the sample covariance matrix in the GLASSO estimator (or another sparse precision matrix estimator). This results in a sparse, cellwise robust precision

matrix estimate. Our approach differs from Tarr et al. (2015) in the selection of the initial covariance estimate. We estimate robust correlations and standard deviations separately to get the robust covariances. The resulting covariance matrix is then always positive semidefinite. This leads to a simplification of the estimator, increases the breakdown point, and speeds up computation substantially.

19.3.1 Robust Covariance Matrix Estimation Based on Pairwise Covariances

Tarr et al. (2015) use the approach of Gnanadesikan and Kettenring (1972) to obtain a robust, pairwise covariance estimate. It is based on the idea that the robust covariance of two random variables X and Y can be computed using a robust variance. For the population covariance **C**ov and the population variance **V**ar, the following identity holds

$$\mathbf{C}\mathrm{ov}(X, Y) = \frac{1}{4\alpha\beta}[\mathbf{V}\mathrm{ar}(\alpha X + \beta Y) - \mathbf{V}\mathrm{ar}(\alpha X - \beta Y)] \tag{19.3}$$

with $\alpha = 1/\sqrt{\mathbf{V}\mathrm{ar}(X)}$, $\beta = 1/\sqrt{\mathbf{V}\mathrm{ar}(Y)}$. If **V**ar is replaced by a robust variance estimator, a robust covariance estimate can be obtained.

This approach has two drawbacks. Firstly, the addition and subtraction of different variables leads to a propagation of the outliers. Therefore, the resulting estimator has a breakdown point of less than 25 % under cellwise contamination. Secondly, the resulting covariance matrix is not necessarily positive semidefinite. Therefore, Tarr et al. (2015) need to apply methods that "make" the matrix positive semidefinite to be able to use this covariance matrix estimate as a replacement of the sample covariance matrix in a sparse precision matrix estimator. To this end, they use the orthogonalized Gnanadesikan-Kettenring (OGK) approach (Maronna and Zamar 2002) as well as the computation of the nearest positive (semi)definite (NPD) matrix as suggested by Higham (2002). Starting from an estimate $\tilde{\mathbf{S}} \in \mathbb{R}^{p\times p}$ for the covariance matrix of the data $\mathbf{X} \in \mathbb{R}^{n\times p}$, NPD finds the closest positive semidefinite matrix $\mathbf{S}$ to the covariance estimate $\tilde{\mathbf{S}}$ in terms of the Frobenius norm

$$\mathbf{S} = \min_{\hat{\mathbf{S}} \succeq 0} \|\tilde{\mathbf{S}} - \hat{\mathbf{S}}\|_F,$$

where $\|\mathbf{A}\|_F = \sum_{j,k=1}^{p} a_{jk}^2$ for a matrix $\mathbf{A} = (a_{jk})_{j,k=1,\ldots,p} \in \mathbb{R}^{p\times p}$ and $\mathbf{A} \succeq 0$ denotes a positive semidefinite matrix. An algorithm to compute the nearest matrix $\mathbf{S}$ is implemented in the R-package `Matrix` under the command `nearPD()`. In our simulations, we observed that NPD gave in general better results than OGK and could also be computed considerably faster.

19.3.2 Robust Covariance Matrix Estimation Based on Pairwise Correlations

In contrast to Tarr et al. (2015), we use a robust correlation estimator $r(\cdot)$ to estimate the pairwise covariance matrix $(s_{jk}) \in \mathbb{R}^{p\times p}$

$$s_{jk} = \text{scale}(\mathbf{x}^j)\,\text{scale}(\mathbf{x}^k) r(\mathbf{x}^j, \mathbf{x}^k) \qquad j,k = 1,\ldots,p \tag{19.4}$$

from the data $\mathbf{X} = (\mathbf{x}^1, \ldots, \mathbf{x}^p) \in \mathbb{R}^{n\times p}$, where scale() is a robust scale estimate like the median absolute deviation or the Q_n-estimator (Rousseeuw and Croux 1993). Both estimators are equally robust with a breakdown point of 50 %. Since the Q_n-estimator is more efficient at the Gaussian model and does not need a location estimate, we opt for this scale estimate. The amount of contamination that the resulting covariance matrix $\mathbf{S} = (s_{jk})_{j,k=1,\ldots,p}$ can withstand depends then on the breakdown point of the scale estimator used (see Sect. 19.5). Using the Q_n-scale, we obtain an estimator with a breakdown point of 50 % under cellwise contamination.

There are different possibilities for choosing a robust correlation estimator. Gaussian rank correlation (e.g., Boudt et al. 2012) is defined as the sample correlation estimated from the Van Der Waerden scores (or normal scores) of the data

$$r_{Gauss}(\mathbf{x}^j, \mathbf{x}^k) = \frac{\sum_{i=1}^n \Phi^{-1}(\frac{R(x_{ij})}{n+1})\Phi^{-1}(\frac{R(x_{ik})}{n+1})}{\sum_{i=1}^n (\Phi^{-1}(\frac{i}{n+1}))^2}, \tag{19.5}$$

where $R(x_{ij})$ denotes the rank of x_{ij} among all elements of $\mathbf{x}^j$, the jth column of the data matrix. Similarly $R(x_{ik})$ stands for the rank of x_{ik} among all elements of $\mathbf{x}^k$. Gaussian rank correlation is robust and consistent at the normal model. Still it is asymptotically equally efficient as the sample correlation coefficient at normal data. This makes it a very appealing robust correlation estimator. Note that the Gaussian rank correlations can easily be computed as the sample covariance matrix from the ranks $R(x_{ij})$ of the data. Since the sample covariance matrix is positive semidefinite, the covariance matrix $\mathbf{S}$ using Gaussian rank correlation is also positive semidefinite. Therefore, we do not need to apply NPD or OGK to obtain a positive semidefinite covariance estimate. This saves computation time and simplifies the final precision matrix estimator.

Another robust correlation estimator is Spearman correlation (Spearman 1904). It is defined as the sample correlation of the ranks of the observations:

$$r_{Spearman}(\mathbf{x}^j, \mathbf{x}^k) = \sum_{i=1}^n \frac{(R(x_{ij}) - \frac{n+1}{2})(R(x_{ik}) - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^n (R(x_{ij}) - \frac{n+1}{2})^2 \sum_{i=1}^n (R(x_{ik}) - \frac{n+1}{2})^2}}.$$

Spearman correlation is slightly less efficient than Gaussian rank correlation. Additionally, it is not consistent at the normal model. To obtain consistency, the

correlation estimator needs to be nonlinearly transformed. The transformation, however, destroys the positive semidefiniteness of the estimator $\mathbf{S}$, and therefore we do not apply it. In our opinion, the inconsistency is not a huge problem because the asymptotic bias of the Spearman correlation is at most 0.018 (Boudt et al. 2012). This is also confirmed by the simulations in Sect. 19.6, where similar results are obtained with Spearman correlation as with Gaussian rank correlation.

We also consider Quadrant correlation (Blomqvist 1950). Quadrant correlation is defined as the frequency of centered observations in the first and third quadrant, minus the frequency of centered observations in the second and fourth quadrant

$$r_{Quadrant}(\mathbf{x}^j, \mathbf{x}^k) = \frac{1}{n}\sum_{i=1}^{n} \text{sign}((x_{ij} - \text{med}_{\ell=1,\ldots,n}\, x_{\ell j})(x_{ik} - \text{med}_{\ell=1,\ldots,n}\, x_{\ell k})),$$

where sign$(\cdot)$ denotes the sign-function. Quadrant correlation is less efficient than Gaussian rank correlation and Spearman correlation (Croux and Dehon 2010). Like Spearman correlation, Quadrant correlation is only consistent at the normal model if a transformation is applied to the correlation estimate. The final covariance matrix of the consistent Quadrant correlation is no longer positive semidefinite. Since we need a positive semidefinite covariance matrix, we opt for the inconsistent Quadrant correlation. Note that the asymptotic bias at the normal distribution of the inconsistent Quadrant correlation is substantially higher than for Spearman correlation. Taking all this drawback of Quadrant correlation into account, it is not a surprise that we obtain worse simulation results with Quadrant correlation than with Spearman or Gaussian rank correlation in Sect. 19.6.

19.3.3 Cellwise Robust Precision Matrix Estimation

To obtain a cellwise robust precision matrix estimator, we adapt the definition of the GLASSO estimator given in (19.2). Recall that GLASSO takes the sample covariance estimator as an input and returns a sparse estimate of the precision matrix as an output. We will replace the sample covariance matrix by the cellwise robust covariance matrices $\mathbf{S}$ of Sects. 19.3.1 and 19.3.2 in order to obtain a cellwise robust, sparse precision matrix estimator. Hence, we obtain the following estimator

$$\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}) = \underset{\substack{\boldsymbol{\Theta}=(\theta_{jk})\in\mathbb{R}^{p\times p}\\ \boldsymbol{\Theta}\succ 0}}{\arg\max}\ \log\det(\boldsymbol{\Theta}) - \text{tr}(\mathbf{S}\boldsymbol{\Theta}) - \rho\sum_{j,k=1}^{p}|\theta_{jk}|, \tag{19.6}$$

If $\mathbf{S}$ is a robust covariance matrix based on pairwise correlations as in Sect. 19.3.2, we refer to $\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X})$ as "correlation based precision matrix estimator." If $\mathbf{S}$ is a covariance matrix based on pairwise covariances as in Sect. 19.3.1, we call $\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X})$ "covariance based precision matrix estimator." Since the algorithm for computing

the GLASSO only requires a positive semidefinite, symmetric matrix $\mathbf{S}$ as an input and not the data, we use it to compute $\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X})$.

Like for the original GLASSO algorithm, the final precision matrix estimate $\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X})$ will always be positive definite as long as the initial covariance matrix $\mathbf{S}$ is positive semidefinite, even if $p > n$. Therefore, it is important that the initial covariance estimate $\mathbf{S}$ is positive semidefinite.

The final precision matrix estimator $\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X})$ will inherit the breakdown point of the initial covariance matrix $\mathbf{S}$ (see Sect. 19.5). As a result, the correlation based precision matrix estimator has a breakdown point of 50 % under cellwise contamination, while the covariance estimators based on pairwise covariances can have a breakdown point of at most 25 % under cellwise contamination.

The covariance matrices based on pairwise correlations we considered (i.e., the matrices based on Gaussian correlation, Spearman correlation, and Quadrant correlation) are all positive semidefinite. Indeed, they can be computed as sample correlation matrices of transformed data. For instance, the quadrant correlation matrix is a sample correlation matrix of the signs of the differences of the observations to their median. In contrast, covariance matrices based on pairwise covariances need to be transformed to be positive semidefinite for which we used the NPD method described in Sect. 19.3.1. Additionally, all pairwise robust covariances need to be computed according to (19.3), which may become very time consuming. Therefore, the correlation based precision matrix estimators are much faster to compute than the covariance based precision matrix estimators.

To sum up, correlation based precision matrix estimators are faster to compute and feature a higher breakdown point under cellwise contamination than covariance based precision matrix estimators.

19.4 Selection of the Regularization Parameter ρ

When selecting the regularization parameter ρ, a good trade-off between a high value of the likelihood function and the sparseness of the final precision matrix has to be found. The two most common methods to find the optimal trade-off are the Bayesian Information Criterion (BIC) and cross validation (CV).

The BIC for a L_1-regularized precision matrix estimator $\hat{\boldsymbol{\Theta}}_\rho$ for a fixed value of ρ has been given in Yuan and Lin (2007):

$$BIC_{classic}(\rho) = -\log\det\hat{\boldsymbol{\Theta}}_\rho + \operatorname{tr}(\hat{\boldsymbol{\Theta}}_\rho\hat{\boldsymbol{\Sigma}}) + \frac{\log n}{n}\sum_{i\leq j}\hat{e}_{ij}(\rho)$$

with $\hat{\boldsymbol{\Sigma}}$ the sample covariance estimate and $\hat{e}_{ij} = 1$ if $(\hat{\boldsymbol{\Theta}}_\rho)_{ij} \neq 0$ and $\hat{e}_{ij} = 0$ otherwise. To obtain a cellwise robust BIC criterion, we replace $\hat{\boldsymbol{\Sigma}}$ by a cellwise

robust covariance matrix $\mathbf{S}$ and use a cellwise robust precision matrix estimator $\hat{\boldsymbol{\Theta}}_\rho$:

$$BIC(\rho) = -\log\det\hat{\boldsymbol{\Theta}}_\rho + \operatorname{tr}(\hat{\boldsymbol{\Theta}}_\rho \mathbf{S}) + \frac{\log n}{n}\sum_{i\le j}\hat{e}_{ij}(\rho).$$

Computing the value of BIC over a grid, the value ρ yielding the lowest BIC is chosen.

To perform K-fold cross validation, the data first has to be split into K blocks of nearly equal size n_k ($k = 1,\ldots,K$). Each block k is left out once and used as test data $(\mathbf{x}^1_{(k)},\ldots,\mathbf{x}^p_{(k)})$. On the remaining data, the precision matrix estimate $\hat{\boldsymbol{\Theta}}_\rho^{(-k)}$ is computed using the regularization parameter ρ. As an evaluation criterion, the negative log-likelihood on the test data is computed

$$L^{(k)}(\rho) = -\log\det\hat{\boldsymbol{\Theta}}_\rho^{(-k)} + \operatorname{tr}(\mathbf{S}^{(k)}\hat{\boldsymbol{\Theta}}_\rho^{(-k)}),$$

where $\mathbf{S}^{(k)}$ is the initial robust covariance estimate computed on the test data, i.e.

$$(\mathbf{S}^{(k)})_{ij} = \operatorname{scale}(\mathbf{x}^i_{(k)})\operatorname{scale}(\mathbf{x}^j_{(k)})r(\mathbf{x}^i_{(k)},\mathbf{x}^j_{(k)}) \qquad i,j = 1,\ldots,p$$

exactly as in Eq. (19.4). By using a robust covariance estimate computed from the test data, outliers present in the test data will not affect the cross-validation criterion too much. This is done over a range of values of ρ. The value of ρ minimizing the negative log-likelihood is chosen as the final regularization parameter

$$\hat{\rho} = \arg\min_\rho \frac{1}{K}\sum_{k=1}^{K} L^{(k)}(\rho). \tag{19.7}$$

As pointed out by a referee, it could occur that some of the test data sets include a percentage of outliers exceeding the breakdown point of the precision matrix estimator, leading to possible breakdown of the cross validation procedure. In our numerical experiments, with contamination levels low compared to the breakdown point and independent for different cells, we did not face this problem. Replacing the sum in (19.7) by a median, for instance, may provide a way out.

To select a grid of values of ρ, we suggest to use the heuristic approach implemented in the `huge`-package (Zhao et al. 2014). It chooses a logarithmic spaced grid of ten values. The largest value of the grid depends on the value of the initial covariance matrix $\mathbf{S}$

$$\rho_{\max} = \max\left(\max_{(i,j)\in\{1,\ldots,p\}^2}(\mathbf{S}-\mathbf{I}_p)_{ij}, -\min_{(i,j)\in\{1,\ldots,p\}^2}(\mathbf{S}-\mathbf{I}_p)_{ij}\right).$$

The smallest value of the grid is then a tenth of the largest value $\rho_{\min} = 0.1\rho_{\max}$. To obtain a logarithmic spaced grid, ten equally spaced values between $\log(\rho_{\min})$

and $\log(\rho_{\max})$ are transformed via the exponential function. We will use this grid of ρ-values in the remainder of the paper.

In general, the BIC criterion can be computed faster than cross validation. However, BIC tends to select too sparse models in practice. In our opinion, the gain in accuracy when using cross validation is worth the increased computation time. Therefore, we will use five-fold cross validation in the remainder of the paper.

19.5 Breakdown Point

In Sect. 19.3, we obtain precision matrix estimators by replacing the sample covariance matrix in the GLASSO with robust covariance matrices. It is not immediately clear if the cellwise robustness of the initial covariance estimator translates to cellwise robustness of the final precision matrix estimator. Theorem 19.1 shows that the final precision matrix estimator $\hat{\boldsymbol{\Theta}}_{\mathbf{S}}$ indeed inherits the breakdown point of the covariance matrix estimator $\mathbf{S}$. Furthermore, we formally show in Proposition 19.1 that the proposed initial covariance matrix estimators based on pairwise correlations are cellwise robust.

One of the most common measurements of robustness is the finite-sample breakdown point. We refer to Maronna et al. (2006) for the standard definition, i.e. under rowwise contamination. The breakdown point denotes the smallest amount of contamination in the data that drives the estimate to the boundary of the parameter space. For example, a location estimator needs to stay bounded, a dispersion estimator needs to stay bounded and away from zero. More formally, define for any symmetric $p \times p$ matrices $\mathbf{A}$ and $\mathbf{B}$

$$D(\mathbf{A}, \mathbf{B}) = \max\{|\lambda_1(\mathbf{A}) - \lambda_1(\mathbf{B})|, |\lambda_p(\mathbf{A})^{-1} - \lambda_p(\mathbf{B})^{-1}|\},$$

where the ordered eigenvalues of a matrix $\mathbf{A}$ are denoted by $0 \leq \lambda_p(\mathbf{A}) \leq \ldots \leq \lambda_1(\mathbf{A})$. We define the *finite-sample breakdown point under cellwise contamination* of a precision matrix estimate $\hat{\boldsymbol{\Theta}}$ as

$$\epsilon_n(\hat{\boldsymbol{\Theta}}, \mathbf{X}) = \min_{m=1,\ldots,n} \{\frac{m}{n} : \sup_{\mathbf{X}^m} D(\hat{\boldsymbol{\Theta}}(\mathbf{X}), \hat{\boldsymbol{\Theta}}(\mathbf{X}^m)) = \infty\}, \tag{19.8}$$

where $\mathbf{X}^m$ denotes a corrupted sample obtained from $\mathbf{X} \in \mathbb{R}^{n \times p}$ by replacing in each column at most m cells by arbitrary values. Similarly, we can define the *explosion* finite-sample breakdown point under cellwise contamination of a covariance matrix estimate $\mathbf{S}$ as

$$\epsilon_n^+(\mathbf{S}, \mathbf{X}) = \min_{m=1,\ldots,n} \{\frac{m}{n} : \sup_{\mathbf{X}^m} |\lambda_1(\mathbf{S}(\mathbf{X})) - \lambda_1(\mathbf{S}(\mathbf{X}^m))| = \infty\}, \tag{19.9}$$

where $\mathbf{X}^m$ denotes a corrupted sample obtained from $\mathbf{X}$ by replacing in each column at most m cells by arbitrary values.

Finally, recall the definition of the explosion breakdown point of a univariate scale estimator scale(·):

$$\epsilon_n^+(\text{scale}, \mathbf{x}) = \min_{m=1,\ldots,n} \{\frac{m}{n} : \sup_{\mathbf{x}^m} \text{scale}(\mathbf{x}^m) = \infty\},$$

where $\mathbf{x}^m$ is obtained from $\mathbf{x} \in \mathbb{R}^n$ by replacing m of the n values by arbitrary values.

To prove the main theorem of this section, we use different properties of eigenvalues, which we summarize in the following lemma.

Lemma 19.1 *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times p}$ *and denote their smallest (largest) eigenvalues by* $\lambda_p(\mathbf{A})$ *(*$\lambda_1(\mathbf{A})$*) and* $\lambda_p(\mathbf{B})$ *(*$\lambda_1(\mathbf{B})$*), respectively. Then the following statements are true:*

(a) If $\mathbf{A}$ *and* $\mathbf{B}$ *are positive semidefinite, then*

$$\lambda_p(\mathbf{AB}) \leq \lambda_1(\mathbf{A})\lambda_p(\mathbf{B}), \tag{19.10}$$

$$\lambda_p(\mathbf{A})\lambda_p(\mathbf{B}) \leq \lambda_p(\mathbf{AB}). \tag{19.11}$$

(b) If $\mathbf{A}$ *and* $\mathbf{B}$ *are symmetric, then*

$$\lambda_1(\mathbf{A} + \mathbf{B}) = \lambda_1(\mathbf{A}) + \lambda_1(\mathbf{B}). \tag{19.12}$$

(c) Denoting $\mathbf{A} = (a_{ij})_{i,j=1,\ldots,p}$*, we have*

$$|\lambda_1(\mathbf{A})| \leq p \max_{i,j=1,\ldots,p} |a_{ij}|. \tag{19.13}$$

Proof The statements of this lemma are proven in Seber (2008). For (a) see 6.76, for (b) see 6.71, for (c) see 6.26a.

Now, we can show that replacing the sample covariance matrix in the GLASSO by a robust covariance matrix $\mathbf{S}$ leads to a precision matrix estimator $\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X})$ that inherits its robustness from $\mathbf{S}$.

Theorem 19.1 *The finite sample breakdown point under cellwise contamination of the robust precision matrix estimator* $\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X})$ *fulfills*

$$\epsilon_n(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}, \mathbf{X}) \geq \epsilon_n^+(\mathbf{S}, \mathbf{X}) \tag{19.14}$$

with $\mathbf{S}$ *a positive semidefinite covariance estimator.*

Proof Let $1 \leq m \leq n$ be the maximum number of cells in a column that have been replaced to arbitrary positions. Since $\mathbf{S}(\mathbf{X}^m)$ is positive semidefinite, $\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m)$ is positive definite (see Banerjee et al. 2008, Theorem 3). The estimate $\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m)$ needs to fulfill the first order condition of (19.6):

$$\mathbf{0} = \hat{\boldsymbol{\Theta}}_{\mathbf{S}}^{-1}(\mathbf{X}^m) - \mathbf{S}(\mathbf{X}^m) - \rho\, \text{Sign}\, \hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m), \tag{19.15}$$

where $(\text{Sign}\,\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m))_{jk} = \text{sign}\,\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m)_{jk}$ for $j,k = 1,\ldots,p$. If $\hat{\boldsymbol{\Theta}}_{\mathbf{S}}$ has zero components, the first order condition (19.15) corresponds to a subdifferential and the sign function at 0 needs to be interpreted as the set $[-1,1]$ (Bertsekas 1995). We then obtain

$$\mathbf{I}_p = (\mathbf{S}(\mathbf{X}^m) + \rho\,\text{Sign}\,\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m))\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m).$$

Thus, the smallest eigenvalue fulfills

$$1 = \lambda_p(\mathbf{I}_p) = \lambda_p((\mathbf{S}(\mathbf{X}^m) + \rho\,\text{Sign}\,\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m))\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m)).$$

Using (19.10), we get

$$1 \leq \lambda_1(\mathbf{S}(\mathbf{X}^m) + \rho\,\text{Sign}\,\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m))\lambda_p(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m)).$$

By definition $\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m)$ is always symmetric, therefore also $\rho\,\text{Sign}(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m))$. As a result, (19.12) yields

$$1 \leq [\lambda_1(\mathbf{S}(\mathbf{X}^m)) + \lambda_1(\rho\,\text{Sign}\,\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m))]\lambda_p(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m)).$$

As $\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m)$ is positive definite, we obtain

$$\frac{1}{\lambda_p(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m))} \leq \lambda_1(\mathbf{S}(\mathbf{X}^m)) + \rho\lambda_1(\text{Sign}\,\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m)).$$

From the definition of the Sign-function, we know that $|(\text{Sign}\,\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m))_{ij}| \leq 1$. Together with (19.13), this yields

$$|\lambda_1(\text{Sign}\,\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m))| \leq p, \tag{19.16}$$

resulting in

$$\lambda_p(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^m))^{-1} \leq \lambda_1(\mathbf{S}(\mathbf{X}^m)) + \rho p. \tag{19.17}$$

From the definition of the explosion breakdown point (19.9), we know that for every $\tilde{m} < n\epsilon_n^+(\mathbf{S},\mathbf{X})$ there exists an $M < \infty$ such that

$$\lambda_1(\mathbf{S}(\mathbf{X}^{\tilde{m}})) \leq M + \lambda_1(\mathbf{S}(\mathbf{X})). \tag{19.18}$$

Using (19.17) in (19.18) yields

$$0 \leq \lambda_p(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^{\tilde{m}}))^{-1} \leq \lambda_1(\mathbf{S}(\mathbf{X}^{\tilde{m}})) + \rho p \leq M + \lambda_1(\mathbf{S}(\mathbf{X})) + \rho p.$$

Together with the triangle inequality this gives

$$|\lambda_p(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^{\tilde{m}}))^{-1} - \lambda_p(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}))^{-1}| \leq \lambda_p(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^{\tilde{m}}))^{-1} + \lambda_p(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}))^{-1} \tag{19.19}$$

$$\leq M + \lambda_1(\mathbf{S}(\mathbf{X})) + \lambda_p(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}))^{-1} + \rho p. \tag{19.20}$$

To obtain a bound for the largest eigenvalue $\lambda_1(\hat{\boldsymbol{\Theta}}_S(\mathbf{X}^m))$, denote for any matrix $\boldsymbol{\Theta} \succ 0$

$$Q(\boldsymbol{\Theta}, \mathbf{X}) = \log\det \boldsymbol{\Theta} - \operatorname{tr}(\mathbf{S}(\mathbf{X})\boldsymbol{\Theta}) - \rho \sum_{j,k=1}^{p} |\theta_{jk}|.$$

For the identity matrix, we obtain for contaminated data $\mathbf{X}^m$

$$Q(\mathbf{I}_p, \mathbf{X}^m) = 0 - \operatorname{tr}(\mathbf{S}(\mathbf{X}^m)) - \rho p \geq -p\lambda_1(\mathbf{S}(\mathbf{X}^m)) - \rho p$$

since $\operatorname{tr}(\mathbf{A}) = \sum_{j=1}^{p} \lambda_j(\mathbf{A}) \leq p\lambda_1(\mathbf{A})$ for any matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$. Using Eq. (19.18), this leads to

$$Q(\mathbf{I}_p, \mathbf{X}^{\tilde{m}}) \geq -pM - p\lambda_1(\mathbf{S}(\mathbf{X})) - \rho p.$$

For any matrix $\tilde{\boldsymbol{\Theta}} \succ 0$, we obtain with (19.11)

$$\operatorname{tr}(\mathbf{S}\tilde{\boldsymbol{\Theta}}) = \sum_{j=1}^{p} \lambda_j(\mathbf{S}\tilde{\boldsymbol{\Theta}}) \geq \lambda_p(\mathbf{S}\tilde{\boldsymbol{\Theta}}) \geq \lambda_p(\mathbf{S})\lambda_p(\tilde{\boldsymbol{\Theta}}) \geq 0. \tag{19.21}$$

Furthermore, (19.13) yields

$$\sum_{i,j=1}^{p} |\tilde{\theta}_{jk}| \geq \max_{j,k=1,\ldots,p} |\tilde{\theta}_{jk}| \geq \frac{1}{p}\lambda_1(\tilde{\boldsymbol{\Theta}}). \tag{19.22}$$

Equations (19.21) and (19.22) lead to

$$Q(\tilde{\boldsymbol{\Theta}}, \mathbf{X}^m) = \log\det \tilde{\boldsymbol{\Theta}} - \operatorname{tr}(S\tilde{\boldsymbol{\Theta}}) - \rho \sum_{j,k=1}^{p} |\theta_{jk}|$$

$$\leq p \log \lambda_1(\tilde{\boldsymbol{\Theta}}) - \frac{\rho}{p}\lambda_1(\tilde{\boldsymbol{\Theta}})$$

because $\det(\mathbf{A}) = \prod_{j=1}^{p} \lambda_j(\mathbf{A}) \leq \lambda_1(\mathbf{A})^p$ for any matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$.

The function $x \mapsto p \log x - \rho x/p$ is concave and attains its maximum at $x = p^2/\rho$. Therefore, there exists a finite constant $M^* > p^2/\rho$, such that

$$p \log M^* - \frac{\rho}{p} M^* = -pM - p\lambda_1(\mathbf{S}(\mathbf{X})) - \rho p.$$

As a results we know that any matrix $\tilde{\boldsymbol{\Theta}}$ with $\lambda_1(\tilde{\boldsymbol{\Theta}}) > M^*$ is not optimizing (19.6) since $Q(\mathbf{I}_p, \mathbf{X}^{\tilde{m}}) > Q(\tilde{\boldsymbol{\Theta}}, \mathbf{X}^{\tilde{m}})$. Hence,

$$0 \leq \lambda_1(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^{\tilde{m}})) \leq M^*.$$

Together with the triangular inequality, this yields

$$|\lambda_1(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^{\tilde{m}})) - \lambda_1(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}))| \leq \lambda_1(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^{\tilde{m}})) + \lambda_1(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X})) \tag{19.23}$$

$$\leq M^* + \lambda_1(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X})). \tag{19.24}$$

Thus, (19.19) and (19.23) lead to

$$\sup_{\mathbf{X}^{\tilde{m}}} D(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}), \hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}^{\tilde{m}})) \leq \max\{M + \lambda_1(\mathbf{S}(\mathbf{X})) + \rho p + \lambda_p(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}))^{-1}, M^* + \lambda_1(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}))\}$$

for any $\tilde{m} < n\epsilon_n^+(\mathbf{S}, \mathbf{X})$, yielding (19.14).

We still need to verify that the covariance matrix estimator based on pairwise correlations has a high explosion breakdown point under cellwise contamination.

Proposition 19.1 *The explosion breakdown point under cellwise contamination of the covariance estimator based on pairwise correlations as defined in (19.4) depends on the explosion breakdown point of the scale estimator used*

$$\epsilon_n^+(\mathbf{S}, \mathbf{X}) \geq \max_{j=1,\ldots,p} \epsilon_n^+(\text{scale}, \mathbf{x}^j). \tag{19.25}$$

Proof Using the triangular inequality, (19.13), (19.4) and the fact that a correlation has an absolute value smaller than 1, we obtain

$$|\lambda_1(\mathbf{S}(\mathbf{X})) - \lambda_1(\mathbf{S}(\mathbf{X}^m))| \leq |\lambda_1(\mathbf{S}(\mathbf{X}))| + p \max_{j,k=1,\ldots,p} |\text{scale}((\mathbf{X}^m)^j)||\text{scale}((\mathbf{X}^m)^k)|$$

for any $m \in \{1, \ldots, n\}$, where $(\mathbf{X}^m)^j$ denotes the jth column of matrix $\mathbf{X}^m$, and therefore (19.25).

Note that the explosion breakdown point of the scale estimator in (19.25) is the breakdown point of a univariate estimator. Breakdown points of scale estimators have been studied extensively (see. e.g., Rousseeuw and Croux 1993). The median

absolute deviation as well as the Q_n-estimator has an explosion breakdown point of 50 %, resulting in a breakdown point of 50 % under cellwise contamination for the correlation based precision matrix estimator proposed in Sect. 19.3.

19.6 Simulations

In this section, we present a simulation study to compare the performance of the estimators introduced in Sect. 19.3. For the correlation based precision matrix estimator, we choose the Q_n-estimator as a scale. As robust correlation, we use Gaussian rank correlation, Spearman correlation, and Quadrant correlation, resulting in the three different estimators "GlassoGaussQn," "GlassoSpearmanQn," and "GlassoQuadQn," respectively. As a point of reference, we also include the non-robust, classical GLASSO (19.2) and abbreviate it as "GlassoClass." Additionally, we compute a covariance based precision matrix estimate, where we choose Q_n as the scale estimator and NPD to obtain a positive semidefinite covariance estimate ("GlassoNPDQn"). This estimator represents the class of estimators studied by Tarr et al. (2015).

To compare to a rowwise, but not cellwise robust estimator that can be computed in high dimensions, we consider the spatial sign covariance matrix (Visuri et al. 2000)

$$\mathbf{S}_{sign}^{incons}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} U(\mathbf{x}_i - \hat{\boldsymbol{\mu}})U(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top, \tag{19.26}$$

where $U(\mathbf{y}) = \|\mathbf{y}\|_2^{-1}\mathbf{y}$ if $\mathbf{y} \neq \mathbf{0}$ and $U(\mathbf{y}) = 0$ otherwise, and $\|\mathbf{y}\|_2$ stands for the Euclidean norm. The location estimator $\hat{\boldsymbol{\mu}}$ is the spatial median, i.e. the minimizer of $\sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\mu}\|_2$. Since only the eigenvectors of (19.26) are consistent estimators for the eigenvectors of the covariance matrix at the normal model, we still need to compute consistent eigenvalues. Let $\mathbf{U}$ denote the matrix of eigenvectors of (19.26). The eigenvalues of the covariance matrix are then given by the marginal variances of $\mathbf{U}^\top\mathbf{x}_1, \ldots, \mathbf{U}^\top\mathbf{x}_n$. To robustly estimate these marginal variances, we use the robust scale estimator Q_n. Denote the matrix of robust eigenvalues as $\boldsymbol{\Lambda} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$. Then the consistent spatial sign covariance matrix is

$$\mathbf{S}_{sign}(\mathbf{X}) = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top.$$

The spatial sign covariance matrix is positive semidefinite. Therefore, we use it as an input in the GLASSO, as in Eq. (19.6), to obtain a sparse precision matrix estimate which is robust under rowwise contamination. We refer to this precision matrix estimator as "GlassoSpSign." Finally, we also add the inverse of the classical sample covariance matrix (19.1) as a benchmark ("Classic"), where it can be computed. For

all estimators, we select the regularization parameter ρ via five-fold cross validation over a logarithmic spaced grid (see Sect. 19.4).

Sampling Schemes We use in total four sampling schemes covering the scenarios of a banded precision matrix, a sparse precision matrix, a dense precision matrix (Cai et al. 2011), and a diagonal precision matrix. Each sampling scheme is defined through the true precision matrix $\boldsymbol{\Theta}_0 \in \mathbb{R}^{p\times p}$ for $i, j = 1, \ldots, p$:

- "banded": $(\boldsymbol{\Theta}_0)_{ij} = 0.6^{|i-j|}$
- "sparse": $\boldsymbol{\Theta}_0 = \mathbf{B} + \delta \mathbf{I}_p$ with $\mathbb{P}[b_{ij} = 0.5] = 0.1$ and $\mathbb{P}[b_{ij} = 0] = 0.9$ for $i \neq j$. The parameter δ is chosen such that the conditional number of $\boldsymbol{\Theta}_0$ equals p. Then the matrix is standardized to have unit diagonals.
- "dense": $(\boldsymbol{\Theta}_0)_{ii} = 1$ and $(\boldsymbol{\Theta}_0)_{ij} = 0.5$ for $i \neq j$
- "diagonal": $(\boldsymbol{\Theta}_0)_{ii} = 1$ and $(\boldsymbol{\Theta}_0)_{ij} = 0$ for $i \neq j$

For each sampling scheme, we generate $M = 100$ samples of size $n = 100$ from a multivariate normal $\mathcal{N}(0, \boldsymbol{\Theta}_0^{-1})$. We take as dimension $p = 60$ and $p = 200$.

Contamination To simulate contamination, we use two different contamination settings (Finegold and Drton 2011): (1) To every generated data set, we add 5 or 10 % of cellwise contamination. Therefore, we randomly select 5 and 10 % of the cells and draw them from a normal $\mathcal{N}(10, 0.2)$. (2) To simulate model deviation, we draw all observations from an alternative t-distribution $t^*_{100,2}(\mathbf{0}, \boldsymbol{\Theta}_0^{-1})$ of dimension 100 with 2 degree of freedom.

Recall that a multivariate t-distributed random variable $\mathbf{x} \sim t_{n,\nu}(\mathbf{0}, \boldsymbol{\Psi})$ is defined as a multivariate normally distributed random variable $\mathbf{y} = (y_1, \ldots, y_p)^\top \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Psi})$ divided by a gamma distributed variable $\tau \sim \Gamma(\nu/2, \nu/2)$

$$\mathbf{x} = \frac{\mathbf{y}}{\sqrt{\tau}}.$$

To obtain an alternative t-distributed random variable $\mathbf{x} = (x_1, \ldots, x_p)^\top \sim t^*_{n,\nu}(\mathbf{0}, \boldsymbol{\Psi})$, we draw p independent divisors $\tau_j \sim \Gamma(\nu/2, \nu/2)$ for the different variables $j = 1, \ldots, p$

$$x_j = \frac{y_j}{\sqrt{\tau_j}}.$$

The heaviness of the tails is then different for different variables of $\mathbf{x}$.

Performance Measures We assess the performance of the estimators using the Kullback-Leibler divergence (Bühlmann and van de Geer 2011, pp. 437)

$$KL(\hat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}_0) = \operatorname{tr}(\boldsymbol{\Theta}_0^{-1}\hat{\boldsymbol{\Theta}}) - \log\det(\boldsymbol{\Theta}_0^{-1}\hat{\boldsymbol{\Theta}}) - p.$$

It measures how close the obtained estimate $\hat{\boldsymbol{\Theta}}$ is to the true parameter $\boldsymbol{\Theta}_0$. Lower values represent a better estimate. If the estimator is equal to the true

precision matrix, the Kullback-Leibler distance is equal to zero. The less accurate the precision matrix is estimated, the higher the value of the Kullback-Leibler distance becomes.

To measure how well the sparseness of the true precision matrix is recovered, we also look at false positive (FP) and false negative (FN) rates:

$$FP = \frac{|\{(i,j) : i = 1,\dots,n; j = 1,\dots,p : (\hat{\boldsymbol{\Theta}})_{ij} \neq 0 \wedge (\boldsymbol{\Theta}_0)_{ij} = 0\}|}{|\{(i,j) : i = 1,\dots,n; j = 1,\dots,p : (\boldsymbol{\Theta}_0)_{ij} = 0\}|}$$

$$FN = \frac{|\{(i,j) : i = 1,\dots,n; j = 1,\dots,p : (\hat{\boldsymbol{\Theta}})_{ij} = 0 \wedge (\boldsymbol{\Theta}_0)_{ij} \neq 0\}|}{|\{(i,j) : i = 1,\dots,n; j = 1,\dots,p : (\boldsymbol{\Theta}_0)_{ij} \neq 0\}|}$$

The false positive rate gives the percentage of zero-elements in the true precision matrix that are wrongly estimated as nonzero. In contrast, the false negative rate gives the percentage of nonzero-elements in the true precision matrix that are wrongly estimated to be zero. Both values are desired to be as small as possible. However, a large false negative rate has a worse impact since it implies that associations between variables are not found and therefore important information is not used. A large false positive rate indicates that unnecessary associations are included, which "only" complicates the model. Note that if $\boldsymbol{\Theta}_0$ does not contain any zero-entries, the false positive rate is not defined. In graphical modeling, a high false negative rate indicates that many non-zero edges that should be included in the estimated graph are missed. This implies that there are conditional independencies assumed which are not supported by the true graph.

Simulation Results Results for $p = 60$ are given in Table 19.1. For clean data in the banded scenario, the classical GLASSO ("GlassoClass") is performing best, achieving lowest values of KL. Only marginally higher values of KL are obtained by the correlation based precision matrix using Gaussian rank correlation ("GlassoGaussQn") and the regularized spatial sign covariance matrix ("GlassoSpSign"). Their good performance can be explained by their high efficiency at the normal model. Even though this data is clean, the inverse of the sample covariance matrix ("Classic") is performing very poorly. This is due to the low precision of the sample covariance matrix for a data set with $p > n/2$. Regularization of the inverse of the sample covariance matrix is solving the problem, as we see from the classical GLASSO. Note that the sample covariance matrix always gives an FN of zero, since the resulting estimate is not sparse, and should therefore not be considered to evaluate the performance of the sample covariance matrix. The correlation based precision matrix using Spearman correlation ("GlassoSpearmanQn") obtains a slightly higher value of KL than "GlassoGaussQn." It probably suffers from its inconsistency. This also explains why the KL of the correlation based precision matrix using Quadrant correlation ("GlassoQuadQn") is so much higher, since the asymptotic bias of the Quadrant correlation is considerably higher than that of Spearman. The performance of the covariance based precision matrix ("GlassoNPDQn") lies in between "GlassoSpearmanQn" and "GlassoQuadQn."

Table 19.1 Simulation results for $n = 100$ and $p = 60$: Kullback-Leibler criterion (KL), false positive rate (FP), and false negative rate (FN) averaged over $M = 100$ simulations reported for seven estimators and four sampling schemes

		Clean			5 % cellwise			10 % cellwise			Alternative t		
		KL	FP	FN	KL	FP	FN	KL	FP	FN	KL	FP	FN
Banded	GlassoClass	8.97		0.70	55.00		0.95	77.11		0.94	143.16		0.98
	GlassoQuadQn	14.96		0.83	19.20		0.86	24.44		0.90	31.10		0.87
	GlassoGaussQn	9.62		0.75	16.91		0.83	23.52		0.88	28.41		0.84
	GlassoSpearmanQn	10.09		0.76	16.32		0.83	22.69		0.87	27.92		0.84
	GlassoNPDQn	11.90		0.85	21.73		0.91	43.59		0.97	37.34		0.92
	Classic	71.54		0.00	49.24		0.00	61.01		0.00	67.84		0.00
	GlassoSpSign	9.53		0.74	53.92		0.96	77.54		0.95	80.41		0.94
Sparse	GlassoClass	5.87	0.23	0.09	63.70	0.02	0.82	88.81	0.04	0.81	140.09	0.00	0.85
	GlassoQuadQn	10.28	0.15	0.38	14.20	0.12	0.47	19.04	0.09	0.56	26.28	0.10	0.45
	GlassoGaussQn	6.34	0.21	0.11	12.25	0.16	0.30	18.39	0.11	0.49	24.09	0.13	0.28
	GlassoSpearmanQn	6.74	0.21	0.13	11.75	0.16	0.27	17.67	0.12	0.43	23.71	0.14	0.26
	GlassoNPDQn	8.25	0.13	0.23	17.85	0.06	0.47	42.14	0.01	0.82	32.73	0.06	0.52
	Classic	71.54	1.00	0.00	49.39	1.00	0.00	66.83	1.00	0.00	62.79	1.00	0.00
	GlassoSpSign	6.35	0.21	0.11	62.36	0.01	0.83	89.37	0.02	0.83	76.37	0.04	0.65
Dense	GlassoClass	4.40		0.92	42.52		0.96	64.95		0.94	128.00		0.98
	GlassoQuadQn	4.65		0.94	7.66		0.95	11.66		0.96	20.72		0.97
	GlassoGaussQn	4.59		0.93	7.59		0.94	11.70		0.96	20.72		0.96
	GlassoSpearmanQn	4.61		0.94	7.59		0.94	11.80		0.96	20.81		0.97
	GlassoNPDQn	5.01		0.96	13.69		0.98	30.98		0.98	28.43		0.98
	Classic	71.54		0.00	39.88		0.00	49.44		0.00	59.08		0.00
	GlassoSpSign	4.62		0.94	41.65		0.97	65.21		0.96	69.46		0.98
Diagonal	GlassoClass	1.31	0.05	0.00	66.11	0.01	0.00	93.48	0.03	0.00	124.03	0.00	0.00
	GlassoQuadQn	1.55	0.04	0.00	4.60	0.03	0.00	8.68	0.03	0.00	17.69	0.02	0.00
	GlassoGaussQn	1.53	0.04	0.00	4.54	0.04	0.00	8.67	0.03	0.00	17.61	0.02	0.00
	GlassoSpearmanQn	1.55	0.04	0.00	4.57	0.04	0.00	8.68	0.03	0.00	17.78	0.02	0.00
	GlassoNPDQn	1.92	0.02	0.00	11.02	0.00	0.00	33.94	0.00	0.00	25.46	0.00	0.00
	Classic	71.54	1.00	0.00	48.41	1.00	0.00	68.67	1.00	0.00	56.26	1.00	0.00
	GlassoSpSign	1.54	0.04	0.00	62.99	0.01	0.00	93.75	0.02	0.00	66.05	0.01	0.00

Under contamination, the relative performance of the different estimators changes. Clearly, the classical GLASSO is not robust, and it achieves the highest values of KL of all estimators. Also the regularized spatial sign covariance matrix does not perform well. This is no surprise since for 5 % of cellwise contamination, already more than 90 % of the observations are expected to be contaminated. Thus, the level of rowwise contamination is too high for "GlassoSpSign" to obtain reliable results. Best performance under contamination is obtained by the correlation based precision matrices using Gaussian rank or Spearman correlation. They give lowest values of KL for all three contamination schemes. Moderately larger values are obtained by "GlassoQuadQn." Of the cellwise robust estimators, the covariance based precision matrix estimator is performing worst under contamination. It

obtains highest values of KL and FN in all three contamination settings. Under 10 % of cellwise contamination the value of KL of "GlassoNPDQn" is nearly double that of "GlassoSpearmanQn."

Looking at the other three sampling schemes "sparse," "dense," and "diagonal," the conclusions are very similar to that of the banded scheme: For clean data "GlassoClass" is doing best, closely followed by "GlassoGaussQn" and "GlassoSpSign." Under contamination "GlassoGaussQn" and "GlassoSpearmanQn" are performing best, while "GlassoNPDQn" gives worst results of all cellwise robust estimators. For the sparse settings "sparse" and "diagonal" we also compare the different values of the FP and FN. In the setting "diagonal" the values are more or less the same for all estimators (apart from the sample covariance matrix which does not give sparse results and therefore has an FP equal to one). In the setting "sparse," differences are more outspoken. The covariance based precision matrix estimator gives an FN of up to double that of "GlassoGaussQn" or "GlassoSpearmanQn," which is not made up by the slightly lower value of FP. In graphical modeling that means that many nonzero edges are missed by "GlassoNPDQn," while they are correctly identified by "GlassoGaussQn" and "GlassoSpearmanQn."

The simulation results for $p = 200$ are given in Table 19.2. Since $p > n$, the sample covariance matrix cannot be inverted anymore and is excluded from the analysis. Overall, the conclusions are similar to $p = 60$. For clean data, the classical GLASSO performs best. Marginally larger values of KL are obtained by "GlassoGaussQn" and "GlassoSpSign." In comparison with $p = 60$, here also "GlassoSpearmanQn" is doing very well for clean data.

For $p = 200$, we see again that under any type of contamination the classical GLASSO and the regularized spatial sign covariance matrix are not reliable any more. In contrast, the cellwise robust correlation based precision matrix estimators achieve very good results, especially in combination with Gaussian rank or Spearman correlation. Their KL and their FN are lowest of all estimators for all settings considered here. The covariance based correlation estimate is considerably less accurate than the correlation based estimates. Under higher amounts of cellwise contamination "GlassoNPDQn" can have a KL of more than four times the value of the correlation based precision matrix estimators. Besides, its FN is higher in all settings considered.

Since in high-dimensional analysis computation time is important for practical usage of the estimators, Table 19.3 gives an overview of the average computation time that the different estimators require. The computation time was comparable throughout the different simulation schemes. Therefore, we only give averages. Note that the reported computation time includes the selection of ρ via 5-fold crossvalidation. For $p = 60$, the correlation based precision matrices, the classical GLASSO and the regularized spatial sign covariance matrix need very similar computation times. This indicates that the GLASSO algorithm takes most of the computation time and that the computation time of the initial covariance matrices is negligible. In contrast, the covariance based precision matrix estimator is nearly four times slower. For $p = 200$, the classical GLASSO and the regularized spatial sign covariance matrix can be computed fastest. But as they are not robust enough, the estimates are very inaccurate. Computation of the correlation based estimators

Table 19.2 Simulation results for $n = 100$ and $p = 200$: Kullback-Leibler criterion (KL), false positive rate (FP), and false negative rate (FN) averaged over $M = 100$ simulations reported for seven estimators and four sampling schemes

		Clean			5 % cellwise			10 % cellwise			Alternative t		
		KL	FP	FN	KL	FP	FN	KL	FP	FN	KL	FP	FN
Banded	GlassoClass	38.32		0.89	187.21		0.98	262.04		0.98	Inf		0.99
	GlassoQuadQn	56.42		0.94	70.67		0.95	86.97		0.97	112.04		0.96
	GlassoGaussQn	40.18		0.91	63.97		0.94	84.67		0.96	103.21		0.94
	GlassoSpearmanQn	41.53		0.90	61.90		0.93	82.92		0.96	101.81		0.93
	GlassoNPDQn	52.42		0.96	102.91		0.98	200.13		0.99	164.86		0.99
	Classic												
	GlassoSpSign	39.68		0.90	189.47		0.98	265.32		0.98	326.08		0.99
Sparse	GlassoClass	46.65	0.10	0.52	220.90	0.02	0.93	302.94	0.02	0.93	Inf	0.00	0.95
	GlassoQuadQn	60.42	0.09	0.72	75.70	0.07	0.77	93.11	0.05	0.81	119.80	0.06	0.78
	GlassoGaussQn	48.45	0.10	0.54	69.70	0.08	0.69	90.69	0.06	0.79	115.29	0.06	0.71
	GlassoSpearmanQn	49.60	0.10	0.56	68.27	0.08	0.67	88.92	0.06	0.76	114.46	0.06	0.70
	GlassoNPDQn	58.64	0.07	0.64	111.15	0.03	0.84	215.95	0.00	0.95	167.81	0.02	0.85
	Classic												
	GlassoSpSign	47.97	0.10	0.54	223.28	0.01	0.93	306.49	0.01	0.94	339.53	0.01	0.93
Dense	GlassoClass	9.70		0.97	137.73		0.98	214.08		0.98	Inf		0.99
	GlassoQuadQn	10.41		0.98	21.12		0.98	35.06		0.98	66.07		0.99
	GlassoGaussQn	10.35		0.98	20.91		0.98	35.06		0.98	65.54		0.99
	GlassoSpearmanQn	10.39		0.98	21.11		0.98	34.92		0.98	65.80		0.99
	GlassoNPDQn	15.27		0.99	65.43		0.99	146.61		0.99	121.14		0.99
	Classic												
	GlassoSpSign	10.53		0.98	140.17		0.99	217.08		0.98	270.10		0.99
Diagonal	GlassoClass	5.41	0.02	0.00	224.15	0.01	0.00	317.07	0.01	0.00	Inf	0.00	0.00
	GlassoQuadQn	6.14	0.02	0.00	17.12	0.01	0.00	30.71	0.01	0.00	61.51	0.01	0.00
	GlassoGaussQn	6.05	0.02	0.00	17.15	0.01	0.00	30.66	0.01	0.00	61.37	0.01	0.00
	GlassoSpearmanQn	6.07	0.02	0.00	17.08	0.01	0.00	30.71	0.01	0.00	61.18	0.01	0.00
	GlassoNPDQn	10.83	0.00	0.00	63.60	0.00	0.00	167.73	0.00	0.00	114.93	0.00	0.00
	Classic												
	GlassoSpSign	6.29	0.01	0.00	225.91	0.01	0.00	320.17	0.01	0.00	265.02	0.00	0.00

Table 19.3 Computation time (in s) for samples of size $n = 100$ (including selection of ρ via five-fold cross-validation) averaged over $M = 100$ simulations and all simulation schemes reported for seven estimators

	$p = 60$	$p = 200$
GlassoClass	5.93	7.69
GlassoQuadQn	6.13	9.12
GlassoGaussQn	6.09	9.15
GlassoSpearmanQn	5.82	9.01
GlassoNPDQn	22.85	216.79
Classic	0.00	0.02
GlassoSpSign	5.73	8.11

is still very fast here. The estimation including the selection of ρ over a grid of ten values takes less than 10 s. In contrast, estimation of the covariance based precision matrix takes more than 20 times longer.

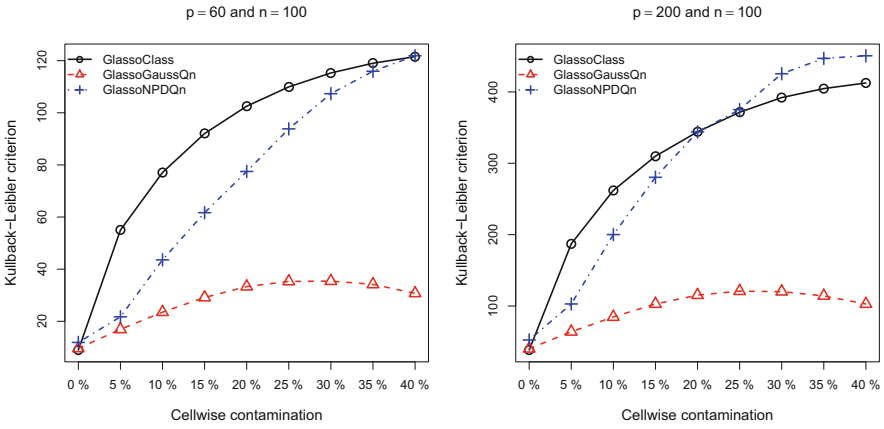

Fig. 19.1 Kullback-Leibler criterion for the "banded" sampling scheme averaged over $M = 100$ simulations reported for various amounts of cellwise contamination and several estimators

Since we advertise the high breakdown point of the correlation based precision matrix estimators, we also look at the performance of the estimators under higher amounts of cellwise contamination, ranging from 0 to 40 %. Figure 19.1 plots the value of KL for the most representative precision matrix estimators for $p = 60$ (left panel) and $p = 200$ (right panel), following the "banded" sampling scheme. As expected, the nonrobust "GlassoClass" results in the highest values of KL. For higher amounts of cellwise contamination, the KL of the "GlassoNPDQn" deteriorates quickly. This is in contrast with the more robust "GlassoGaussQn," where the KL measure remains limited for higher contamination levels, both for $p = 60$ and $p = 200$. The results for the sampling schemes "sparse," "dense," and "diagonal" are comparable to Fig. 19.1 and are therefore omitted.

To summarize, for clean data the classical GLASSO performs best. Under cellwise contamination, "GlassoGaussQn" and "GlassoSpearmanQn" achieve best results. All three estimators can be computed equally fast. Since the "GlassoGaussQn" is consistent and performs similarly well as the classical GLASSO for clean data, we advise the "GlassoGaussQn" for high-dimensional sparse precision matrix estimation under cellwise contamination.

19.7 Applications

In this paper, we describe how a cellwise robust, sparse precision matrix estimator can be obtained. To show the applicability of the introduced estimator to a real-world data set, we use the dataset `stockdata`, which is publicly available through the R-package `huge` (Zhao et al. 2014). It consists of the closing prices of $p = 452$ stocks in the S&P on all trading days between January 1,2003 and January 1, 2008, leading to $n = 1258$ observations. We use the same data transformations

and parameter choices as in Zhao et al. (2012). The estimated graphical models returned by "GlassoClass" and "GlassoGaussQn" are visualized in Panel (a) and (b) of Fig. 19.2. From the plots, we can conclude that the two graphs are very similar. Indeed, only around 2 % of the selected edges in "GlassoClass" are not selected in "GlassoGaussQn," while the percentage is even smaller vice versa. As a result, we assume that `stockdata` is a rather clean data set.

To see how the estimators behave under contamination, we randomly select 5 % of the cells of the data matrix and replace them by replicates of the normal distribution $\mathcal{N}(10, 0.2)$. The graphs estimated by "GlassoClass" and "GlassoGaussQn" from the contaminated data are shown in Panels (c) and (d) of Fig. 19.2, respectively. While the graph estimated by "GlassoGaussQn" hardly differs from the uncontaminated case, "GlassoClass" estimates a graph without any edges. Thus, "GlassoGaussQn" is robust in the sense that the estimate on the contaminated data resembles that of the clean data. In contrast, the nonrobust "GlassoClass" returns a not reliable estimate in the presence of cellwise contaminated data.

Estimating a cellwise robust, sparse precision matrix is not only interesting in graphical models. As an example consider linear discriminant analysis, where each observation belongs to one of K groups. The goal is then to assign a new observation $\mathbf{x} \in \mathbb{R}^p$ to one of those K groups. Assuming a normal distributions $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ for observations of group $k \in \{1, \ldots, K\}$, the Bayes optimal solution is found via the linear discriminant function

$$\delta_k(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log \pi_k,$$

where π_k is the a priori probability of belonging to group k. Replacing Σ^{-1} with the correlation based precision matrix estimated from the centered data (where each observation is centered by the coordinatewise median computed over the observations belonging to the same group) results in a cellwise robust estimator for high-dimensional linear discriminant analysis. The final estimate may not be sparse anymore, but it is very robust under cellwise contamination. Furthermore, it can be computed even if $p > n$.

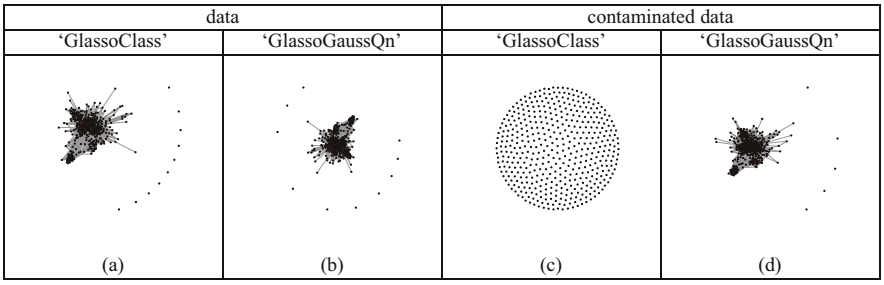

Fig. 19.2 Graphical models estimated from `stockdata`. Every node in the graph corresponds to one of the $p = 452$ stocks

Cellwise robust, sparse precision matrix estimation can also be used to obtain cellwise robust, sparse regression of $\mathbf{y} \in \mathbb{R}^n$ on $\mathbf{X} \in \mathbb{R}^{n \times p}$. Partitioning the joint sample covariance estimate of $(\mathbf{X}, \mathbf{y})$ and its inverse into

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{\mathbf{XX}} & \hat{\boldsymbol{\sigma}}_{\mathbf{Xy}} \\ \hat{\boldsymbol{\sigma}}_{\mathbf{Xy}}^{\top} & \hat{\sigma}_{\mathbf{yy}} \end{pmatrix} \qquad \hat{\boldsymbol{\Theta}} = \begin{pmatrix} \hat{\boldsymbol{\Theta}}_{\mathbf{XX}} & \hat{\boldsymbol{\theta}}_{\mathbf{Xy}} \\ \hat{\boldsymbol{\theta}}_{\mathbf{Xy}}^{\top} & \hat{\theta}_{\mathbf{yy}} \end{pmatrix}$$

the least squares estimator can be rewritten as

$$\hat{\boldsymbol{\beta}}_{LS} = \hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1} \hat{\boldsymbol{\sigma}}_{\mathbf{Xy}} = -\frac{1}{\theta_{yy}} \hat{\boldsymbol{\Theta}}_{\mathbf{Xy}}$$

using the partitioned inverse formula (Seber 2008, Lemma 14.11). With the correlation based precision matrix estimate $\hat{\boldsymbol{\Theta}}_{\mathbf{S}}((\mathbf{X}, \mathbf{y}))$ computed jointly from $(\mathbf{X}, \mathbf{y})$, we obtain a cellwise robust, sparse regression estimate computable in high dimensions

$$\hat{\boldsymbol{\beta}} = -\frac{1}{(\hat{\boldsymbol{\Theta}}_{\mathbf{S}}((\mathbf{X}, \mathbf{y})))_{p+1,p+1}} (\hat{\boldsymbol{\Theta}}_{\mathbf{S}}((\mathbf{X}, \mathbf{y})))_{1:p,p+1}.$$

19.8 Conclusions

We have introduced a cellwise robust, correlation based precision matrix estimator. We put forward the following simple procedure: (1) compute the robust scale estimators Q_n for each variable (2) compute the robust correlation matrix from the normal scores, as in Eq. (19.5) (3) construct then the robust covariance matrix from these correlations and robust scale, as in Eq. (19.4) (4) use the latter as input for the GLASSO, returning $\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X})$. It is formally shown that the proposed estimator features a very high breakdown point under cellwise contamination. As its definition is very simple, the estimator can be computed very fast, even in high-dimensions.

The simulation results presented in Sect. 19.6 discuss the results of the various estimators *including* their selection of the regularization parameter ρ. As can be seen from a small simulation study with $p = 60$, kindly provided by a referee, the bad performance of "GlassoNPDQn" needs to be mainly attributed to the selection of ρ. When "GlassoNPDQn" is run with the regularization parameter estimated by "GlassoGaussQn," the two methods performed similar. This problem also occurred for clean data. Replacing CV by BIC did not help to improve "GlassoNPDQn": The performance in comparison with "GlassoGaussQn" was still similar as in Tables 19.1 and 19.2. Analyzing the reason for the bad performance of "GlassoNPDQn" with respect to the selection of the regularization parameter is left for future research.

Compared to the covariance based approach, a correlation based approach results in a simpler estimator. More importantly, it achieves a substantially higher

breakdown point, is considerably faster to compute, and yields more accurate estimates when the regularization parameter is selected using BIC or the new cross-validation criterion presented in Sect. 19.4.

Acknowledgements The authors wish to acknowledge the helpful comments from two anonymous referees. The first author gratefully acknowledges support from the GOA/12/014 project of the Research Fund KU Leuven.

References

Agostinelli, C., Leung, A., Yohai, V.J., Zamar, R.H.: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. To appear in TEST. http://arxiv.org/abs/1406.6031 (2015)

Alqallaf, F.A., Konis, K.P., Martin, R.D., Zamar, R.H.: Scalable robust covariance and correlation estimates for data mining. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 14–23. ACM, New York (2002)

Alqallaf, F.A., Van Aelst, S., Yohai, V.J., Zamar, R.H.: Propagation of outliers in multivariate data. Ann. Stat. **37**(1), 311–331 (2009)

Banerjee, O., d'Aspremont, A., El Ghaoui, L.: Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J. Mach. Learn. Res. **9**, 485–516 (2008)

Bertsekas, B.: Nonlinear Programming. Athena Scientific, Belmont (1995)

Blomqvist, N.: On a measure of dependence between two random variables. Ann. Math. Stat. **21**(4), 593–600 (1950)

Boudt, K., Cornelissen, J., Croux, C.: The Gaussian rank correlation estimator: robustness properties. Stat. Comput. **22**(2), 471–483 (2012)

Bühlmann, P., van de Geer, S.: Statistics for High-Dimensional Data. Springer, Heidelberg (2011)

Cai, T.T., Liu, W., Luo, X.: A constrained l_1 minimization approach to sparse precision matrix estimation. J. Am. Stat. Assoc. **106**(494), 594–607 (2011)

Cai, T.T., Liu, W., Luo, X.: clime: constrained L1-minimization for Inverse (covariance) matrix estimation. http://CRAN.R-project.org/package=clime (2012) [R package version 0.4.1]

Croux, C., Dehon, C.: Influence functions of the Spearman and Kendall correlation measures. Stat. Methods Appl. **19**(4), 497–515 (2010)

Finegold, M.A., Drton, M.: Robust graphical modeling of gene networks using classical and alternative t-distributions. Ann. Appl. Stat. **5**(2A), 1057–1080 (2011)

Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (2008)

Friedman, J., Hastie, T., Tibshirani, R.: glasso: Graphical lasso estimation of Gaussian graphical models. http://CRAN.R-project.org/package=glasso (2014) [R package version 1.8]

Gnanadesikan, R., Kettenring, J.R.: Robust estimates, residuals and outlier detection with multiresponse data. Biometrics **28**(1), 81–124 (1972)

Higham, N.J.: Computing the nearest correlation matrix - a problem from finance. IMA J. Numer. Anal. **22**(3), 329–343 (2002)

Hsieh, C.J., Sustik, M.A., Dhillon, I.S., Ravikumar, P.: Sparse inverse covariance matrix estimation using quadratic approximation. In: Advances in Neural Information Processing Systems, vol. 24, pp. 2330–2338 (2011). http://nips.cc/

Maronna, R.A., Zamar, R.H.: Robust estimation of location and dispersion for high-dimensional datasets. Technometrics **44**(4), 307–317 (2002)

Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust Statistics, 2nd edn. Wiley, Hoboken (2006)

Ollila, E., Tyler, D.E.: Regularized M-estimators of scatter matrix. IEEE Trans. Signal Process. **62**(22), 6059–6070 (2014)

Ollila, E., Oja, H., Croux, C.: The affine equivariant sign covariance matrix: asymptotic behavior and efficiencies. J. Multivar. Anal. **87**(2), 328–355 (2003)
Rousseeuw, P.J., Croux, C.: Alternatives to the median absolute deviation. J. Am. Stat. Assoc. **88**(424), 1273–1283 (1993)
Rousseeuw, P.J., Van Driessen, K.: A fast algorithm for the minimum covariance determinant estimator. Technometrics **41**(3), 212–223 (1999)
Seber, G.A.F.: A Matrix Handbook for Statisticians. Wiley, Hoboken (2008)
Spearman, C.: The proof and measurement of association between two things. Am. J. Psychol. **15**(1), 72–101 (1904)
Tarr, G.: Quantile based estimation of scale and dependence. PhD thesis, University of Sydney (2014)
Tarr, G., Müller, S., Weber, N.C.: Robust estimation of precision matrices under cellwise contamination. Comput. Stat. Data Anal. (2015). doi:10.1016/j.csda.2015.02.005
Van Aelst, S., Vandervieren, E., Willems, G.: Stahel-donoho estimators with cellwise weights. J. Stat. Comput. Simul. **81**(1), 1–27 (2011)
Visuri, S., Koivunen, V., Oja, H.: Sign and rank covariance matrices. J. Stat. Plann. Inference **91**(2), 557–575 (2000)
Yuan, M., Lin, Y.: Model selection and estimation in the gaussian graphical model. Biometrika **94**(1), 19–35 (2007)
Zhao, T., Liu, H., Roeder, K., Lafferty, J., Wasserman, L.: The `huge` package for high-dimensional undirected graph estimation in R. J. Mach. Learn. Res. **13**, 1059–1062 (2012)
Zhao, T., Liu, H., Roeder, K., Lafferty, J., Wasserman, L.: huge: high-dimensional undirected graph estimation. http://CRAN.R-project.org/package=huge (2014) [R package version 1.2.6]

Chapter 20
Paired Sample Tests in Infinite Dimensional Spaces

Anirvan Chakraborty and Probal Chaudhuri

Abstract The sign and the signed-rank tests for univariate data are perhaps the most popular nonparametric competitors of the t test for paired sample problems. These tests have been extended in various ways for multivariate data in finite dimensional spaces. These extensions include tests based on spatial signs and signed ranks, which have been studied extensively by Hannu Oja and his coauthors. They showed that these tests are asymptotically more powerful than Hotelling's T^2 test under several heavy tailed distributions. In this paper, we consider paired sample tests for data in infinite dimensional spaces based on notions of spatial sign and spatial signed rank in such spaces. We derive their asymptotic distributions under the null hypothesis and under sequences of shrinking location shift alternatives. We compare these tests with some mean based tests for infinite dimensional paired sample data. We show that for shrinking location shift alternatives, the proposed tests are asymptotically more powerful than the mean based tests for some heavy tailed distributions and even for some Gaussian distributions in infinite dimensional spaces. We also investigate the performance of different tests using some simulated data.

Keywords Contaminated data • Gâteaux derivative • Smooth Banach space • Spatial sign • Spatial signed rank • t process

20.1 Introduction

For univariate data, two nonparametric competitors of the t test for one sample and paired sample problems are the sign test and the Wilcoxon signed-rank test. It is well known that these two tests enjoy certain robustness properties, and are more powerful than the t test when the underlying distribution has heavier tails than the Gaussian distribution. The sign and the signed-rank tests have been extended in several ways for data in $\mathbb{R}^d$. Puri and Sen (1971) considered extensions based on

A. Chakraborty (✉) • P. Chaudhuri
Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203, B. T. Road, Kolkata - 700108, India
e-mail: vanchak@gmail.com; probal@isical.ac.in

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_20

coordinatewise signs and ranks. Randles (1989) and Peters and Randles (1990) studied extensions of the sign test and the signed-rank test, respectively, using the notion of interdirections. Chaudhuri and Sengupta (1993) proposed a class of multivariate extensions of the sign test based on data-driven transformations of the sample observations. Hallin and Paindaveine (2002) considered multivariate signed-rank type tests based on interdirections and the ranks of the sample observations computed using pseudo-Mahalanobis distances. Hettmansperger et al. (1994) and Hettmansperger et al. (1997) considered multivariate versions of sign and signed-rank tests using simplices (see also Oja 1999). Möttönen and Oja (1995), Chakraborty et al. (1998), and Marden (1999) used spatial signs and signed ranks (see also Oja 2010) to construct extensions of sign and signed-rank tests for multivariate data. Möttönen et al. (1997) and Chakraborty et al. (1998) studied the asymptotic efficiency of spatial sign and signed-rank tests relative to Hotelling's T^2 test. They showed that the tests based on spatial signs and signed ranks are asymptotically more powerful than the T^2 test under heavy tailed distributions like multivariate t distributions. Further, while Hotelling's T^2 test is optimal for multivariate Gaussian distributions, it was shown that as the data dimension increases, the performance of the spatial sign and the spatial signed-rank tests become closer to Hotelling's T^2 test under multivariate Gaussian distributions.

Nowadays, we often come across data, which are curves or functions observed over an interval, and are popularly known as functional data. Such data are very different from multivariate data in finite dimensional spaces because the data dimension is much larger than the sample size, and also due to the fact that different sample observations may be observed at different sets of points in the interval. However, this type of data can be conveniently handled by viewing them as a sample in some infinite dimensional space, e.g., the space of real-valued functions defined over an interval. Many of the above-mentioned tests cannot be used to analyze such data. This is because the definitions of some of them involve hyperplanes, simplices, etc. constructed using the data, and thus these tests require the data dimension to be smaller than the sample size. Some of the other tests involve inverses of covariance matrices computed from the sample, and such empirical covariance matrices are singular, when the data dimension is larger than the sample size.

Many of the function spaces, where functional data lie, are infinite dimensional Banach spaces. In this paper, we investigate a sign and a signed-rank type test for paired sample problems based on notions of spatial sign and spatial signed rank in such spaces. We derive the asymptotic distributions of the proposed spatial sign and signed-rank statistics under the null hypothesis as well as under suitable sequences of shrinking location shift alternatives. We also compare the asymptotic powers of these tests with some of the paired-sample mean based tests for infinite dimensional data. It is found that these tests outperform the mean based competitors, when the underlying distribution has heavy tails and also for some Gaussian distributions of the data. Some simulation studies are carried out to demonstrate the performance of different tests.

20.2 Sign and Signed-Rank Type Tests

The construction and the study of the paired sample sign and signed-rank type statistics for data in Banach spaces will require several concepts and tools from functional analysis and probability theory in Banach spaces. Let us first recall that for a random variable X with a continuous distribution function F, the "centered rank" of $x \in \mathbb{R}$ with respect to F is given by $2F(x) - 1 = E\{sign(x - X)\}$, where $sign(x)$ is the usual sign function evaluated at x (see Serfling 2004). Note that for any non-zero x, we have $sign(x) = x/|x|$, which is the derivative of the function $x \mapsto |x|$. We will extend the notion of sign into a general Banach space (finite or infinite dimensional) using the concept of a derivative of the norm function on that Banach space. Let $\mathscr{X}$ be a Banach space with norm $||\cdot||$. Let us denote its dual space by $\mathscr{X}^*$, which is the Banach space of real-valued continuous linear functions defined over $\mathscr{X}$.

Definition 20.1 The norm in $\mathscr{X}$ is said to be Gâteaux differentiable at a nonzero $\mathbf{x} \in \mathscr{X}$ with derivative, say, $\mathbf{S}_\mathbf{x} \in \mathscr{X}^*$ if $\lim_{t\to 0}\, t^{-1}(||\mathbf{x} + t\mathbf{h}|| - ||\mathbf{x}||) = \mathbf{S}_\mathbf{x}(\mathbf{h})$ for all $\mathbf{h} \in \mathscr{X}$. A Banach space $\mathscr{X}$ is said to be smooth if the norm in $\mathscr{X}$ is Gâteaux differentiable at every nonzero $\mathbf{x} \in \mathscr{X}$.

As a convention, we take $\mathbf{S}_\mathbf{x} = \mathbf{0}$ if $\mathbf{x} = \mathbf{0}$. Observe that the right-hand limit $\lim_{t\downarrow 0}\, t^{-1}(||\mathbf{x}+t\mathbf{h}||-||\mathbf{x}||)$ always exists by the convexity of the norm function. So, Gâteaux differentiability of the norm function at $\mathbf{x}$ only requires the existence of the left-hand limit. Note that if $\mathscr{X} = \mathbb{R}$, we have $S_x = sign(x) = x/|x|$. Also, if $\mathscr{X} = \mathbb{R}^d$ equipped with the Euclidean norm, it follows that $\mathbf{S}_\mathbf{x} = \mathbf{x}/||\mathbf{x}||$, which is the gradient vector, i.e., the vector of partial derivatives of the Euclidean norm function $||\mathbf{x}|| = (\mathbf{x}^\top\mathbf{x})^{1/2}$ evaluated at $\mathbf{x}$. It is known that Hilbert spaces and L_p spaces for $p \in (1, \infty)$ are smooth. For a Hilbert space $\mathscr{X}$, we have $\mathbf{S}_\mathbf{x} = \mathbf{x}/||\mathbf{x}||$. If $\mathscr{X}$ is an L_p space for some $p \in (1, \infty)$, then $\mathbf{S}_\mathbf{x}(\mathbf{h}) = \int sign\{\mathbf{x}(\mathbf{s})\}|\mathbf{x}(\mathbf{s})|^{p-1}\mathbf{h}(\mathbf{s})d\mathbf{s}/||\mathbf{x}||^{p-1}$ for all $\mathbf{h} \in \mathscr{X}$. The spatial sign of $\mathbf{x}$ in a Banach space $\mathscr{X}$ is the Gâteaux derivative $\mathbf{S}_\mathbf{x}$ as defined above provided it exists.

Definition 20.2 A random element $\mathbf{X}$ in the Banach space $\mathscr{X}$ is said to be Bochner integrable if there exists a sequence $\{\mathbf{X}_n\}_{n\geq 1}$ of simple functions in $\mathscr{X}$ such that $\mathbf{X}_n \to \mathbf{X}$ *almost surely* and $E(||\mathbf{X}_n - \mathbf{X}||) \to 0$ as $n \to \infty$.

It is known that the Bochner expectation of $\mathbf{X}$ exists if $E(||\mathbf{X}||) < \infty$. We refer to Chapters 4 and 5 in Borwein and Vanderwerff (2010) and Section 2 in Chapter 3 of Araujo and Giné (1980) for more details. Henceforth, the expectation of any Banach space valued random element will be in the Bochner sense.

Definition 20.3 Let $\mathscr{X}$ be a smooth Banach space. The spatial rank of $\mathbf{x} \in \mathscr{X}$ with respect to the distribution of a random element $\mathbf{X} \in \mathscr{X}$ is defined to be $\boldsymbol{\Psi}_\mathbf{x} = E(\mathbf{S}_{\mathbf{x}-\mathbf{X}})$.

Note that if $\mathscr{X} = \mathbb{R}$, then $\Psi_x = 2F(x) - 1$ for a continuous distribution F. Further, if $\mathscr{X} = \mathbb{R}^d$ equipped with the Euclidean norm, then $\boldsymbol{\Psi}_\mathbf{x} = E\{(\mathbf{x} - \mathbf{X})/||\mathbf{x} - \mathbf{X}||\}$,

which is the spatial rank of $\mathbf{x}$ with respect to the probability distribution of the random vector $\mathbf{X} \in \mathbb{R}^d$ (see Oja 2010).

Let $(\mathbf{X}_1, \mathbf{Y}_1), (\mathbf{X}_2, \mathbf{Y}_2) \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ be i.i.d. paired observations, where the $\mathbf{X}_i$'s and the $\mathbf{Y}_i$'s take values in a smooth Banach space $\mathscr{X}$. Let $\mathbf{W}_i = \mathbf{Y}_i - \mathbf{X}_i$, $1 \leq i \leq n$, and define $\boldsymbol{\nu} = E(\mathbf{S}_{\mathbf{W}_1})$. A paired sample sign statistic using spatial signs for testing the hypothesis $H_0^{(1)} : \boldsymbol{\nu} = \mathbf{0}$ against $H_1^{(1)} : \boldsymbol{\nu} \neq \mathbf{0}$ is defined as

$$\mathbf{T}_S = n^{-1} \sum_{1 \leq i \leq n} \mathbf{S}_{\mathbf{W}_i}.$$

We reject $H_0^{(1)}$ when $||\mathbf{T}_S||$ is large. Next, define $\boldsymbol{\theta} = E(\mathbf{S}_{\mathbf{W}_1 + \mathbf{W}_2})$. A paired sample signed-rank statistic using spatial signed ranks for testing the hypothesis $H_0^{(2)} : \boldsymbol{\theta} = \mathbf{0}$ against $H_1^{(2)} : \boldsymbol{\theta} \neq \mathbf{0}$ is given by

$$\mathbf{T}_{SR} = 2\{n(n-1)\}^{-1} \sum_{1 \leq i < j \leq n} \mathbf{S}_{\mathbf{W}_i + \mathbf{W}_j}.$$

We reject $H_0^{(2)}$ for large values of $||\mathbf{T}_{SR}||$. In finite dimensional Euclidean spaces, Oja (2010, pp. 83–84) considered the spatial signed-rank statistic

$$\tilde{\mathbf{T}}_{SR} = 2\{n(n-1)\}^{-1} \sum_{1 \leq i < j \leq n} \{\mathbf{S}_{\mathbf{W}_i + \mathbf{W}_j} + \mathbf{S}_{\mathbf{W}_i - \mathbf{W}_j}\},$$

which is derived from the spatial signed-rank function in $\mathbb{R}^d$ for testing the null hypothesis $H_0^{(2)}$ against $H_1^{(2)}$. Further, it is shown in Oja (2010, pp. 83–84) that $\tilde{\mathbf{T}}_{SR}$ and $\mathbf{T}_{SR}$ have the same asymptotic distributions. In any separable Banach space, the asymptotic distributions of $\mathbf{T}_{SR}$ and $\tilde{\mathbf{T}}_{SR}$ are the same. For proving this we first note that the asymptotic distribution of the V-statistic $\hat{\mathbf{T}}_{SR} = n^{-2} \sum_{1 \leq i,j \leq n} \{\mathbf{S}_{\mathbf{W}_i + \mathbf{W}_j} + \mathbf{S}_{\mathbf{W}_i - \mathbf{W}_j}\}$ is the same as that of the corresponding U-statistic $\tilde{\mathbf{T}}_{SR}$, and this can be shown using the techniques used to prove the asymptotic distributions of Banach space valued U-statistics in Borovskikh (1991) (see also the proofs of Theorems 20.1 and 20.2 given later). Since $\sum_{1 \leq i,j \leq n} \mathbf{S}_{\mathbf{W}_i - \mathbf{W}_j} = \mathbf{0}$, it follows that $\hat{\mathbf{T}}_{SR} = n^{-2} \sum_{1 \leq i,j \leq n} \mathbf{S}_{\mathbf{W}_i + \mathbf{W}_j}$. By the same argument as above, this V-statistic has the same asymptotic distribution as that of the corresponding U-statistic $\mathbf{T}_{SR}$. Combining these facts, it follows that the asymptotic distributions of $\mathbf{T}_{SR}$ and $\tilde{\mathbf{T}}_{SR}$ are the same. Recently, a two sample Wilcoxon–Mann–Whitney type test based on spatial ranks in infinite dimensional spaces has been studied by Chakraborty and Chaudhuri (2015). Since $S_x = sign(x)$ if $\mathscr{X} = \mathbb{R}$, it follows that $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ reduce to the univariate sign and signed-rank statistics in this case. Moreover, if $\mathscr{X} = \mathbb{R}^d$, then $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ are the spatial sign and signed-rank statistics for finite dimensional multivariate data studied by Möttönen and Oja (1995), Möttönen et al. (1997) and Marden (1999). Note that the hypothesis $H_0^{(1)} : \boldsymbol{\nu} = \mathbf{0}$ (respectively, $H_0^{(2)} : \boldsymbol{\theta} = \mathbf{0}$) is equivalent to the hypothesis that the spatial median of $\mathbf{W}_1$ (respectively, $\mathbf{W}_1 + \mathbf{W}_2$)

is zero. Suppose that $\mathbf{Y}_1 - \mathbf{X}_1$ has a symmetric distribution about some $\boldsymbol{\eta} \in \mathscr{X}$, i.e., the distribution of $\mathbf{Y}_1 - \mathbf{X}_1 - \boldsymbol{\eta}$ and $\boldsymbol{\eta} - \mathbf{Y}_1 + \mathbf{X}_1$ are the same. Then, it follows that both of $H_0^{(1)}$ and $H_0^{(2)}$ become the hypothesis $\boldsymbol{\eta} = \mathbf{0}$. This holds, in particular, if $\mathbf{X}_1$ and $\mathbf{Y}_1$ are exchangeable, i.e., the distributions of $(\mathbf{X}_1, \mathbf{Y}_1)$ and $(\mathbf{Y}_1, \mathbf{X}_1)$ are the same. Further, if the distribution of $\mathbf{Y}_1 - \mathbf{X}_1$ is symmetric and its mean exists, then both of $H_0^{(1)}$ and $H_0^{(2)}$ are equivalent to the hypothesis $E(\mathbf{Y}_1 - \mathbf{X}_1) = \mathbf{0}$.

20.2.1 Asymptotic Distributions and the Implementation of the Tests

In order to study the asymptotic distributions of $\mathbf{T}_S$ and $\mathbf{T}_{SR}$, we introduce the following definitions.

Definition 20.4 A Banach space $\mathscr{X}$ is said to be of type 2 if there exists a constant $b > 0$ such that for any $m \geq 1$ and independent zero mean random elements $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_m$ in $\mathscr{X}$ with $E(||\mathbf{U}_i||^2) < \infty$, $1 \leq i \leq m$, we have $E(||\sum_{i=1}^m \mathbf{U}_i||^2) \leq b \sum_{i=1}^m E(||\mathbf{U}_i||^2)$.

Definition 20.5 A Banach space $\mathscr{X}$ is said to be p-uniformly smooth for some $p \in (1, 2]$ if for every $q \geq 1$ there exists a constant $\alpha_q > 0$ such that for any zero mean martingale sequence $(\mathbf{M}_m, \mathscr{G}_m)_{m \geq 1}$ in $\mathscr{X}$, we have $E(||\mathbf{M}_m||^q) \leq \alpha_q \sum_{i=1}^m E(||\mathbf{M}_i - \mathbf{M}_{i-1}||^p)^{q/p}$. Here, the sequence $(\mathbf{M}_m)_{m \geq 1}$ is adapted to the filtration $(\mathscr{G}_m)_{m \geq 1}$.

Any 2-uniformly smooth Banach space is of type 2. Hilbert spaces are 2-uniformly smooth, and L_p spaces are $\tilde{p}$-uniformly smooth, where $\tilde{p} = \min(p, 2)$ for $p \in (1, \infty)$.

Definition 20.6 A continuous linear operator $\mathbf{C} : \mathscr{X}^* \to \mathscr{X}$ is said to be symmetric if $\mathbf{y}\{\mathbf{C}(\mathbf{x})\} = \mathbf{x}\{\mathbf{C}(\mathbf{y})\}$ for all $\mathbf{x}, \mathbf{y} \in \mathscr{X}^*$. It is said to be positive if $\mathbf{x}\{\mathbf{C}(\mathbf{x})\} > 0$ for all $\mathbf{x} \in \mathscr{X}^*$.

Definition 20.7 A random element $\mathbf{Z}$ in a separable Banach space $\mathscr{X}$ is said to have a Gaussian distribution with mean $\mathbf{m} \in \mathscr{X}$ and covariance $\mathbf{C}$, which we denote by $G(\mathbf{m}, \mathbf{C})$, if for any $\mathbf{u} \in \mathscr{X}^*$, $\mathbf{u}(\mathbf{Z})$ has a Gaussian distribution on $\mathbb{R}$ with mean $\mathbf{u}(\mathbf{m})$ and variance $\mathbf{u}\{\mathbf{C}(\mathbf{u})\}$. Here, $\mathbf{C} : \mathscr{X}^* \to \mathscr{X}$ is a symmetric positive continuous linear operator.

We refer to Section 7 of Chapter 3 in Araujo and Giné (1980), Borovskikh (1991) and Section 2.4 in Chapter IV of Vakhania et al. (1987) for further details. Define $\boldsymbol{\Pi}_1 : \mathscr{X}^{**} \to \mathscr{X}^*$ and $\boldsymbol{\Pi}_2 : \mathscr{X}^{**} \to \mathscr{X}^*$ as

$$\boldsymbol{\Pi}_1(\mathbf{f}) = E[\mathbf{f}(\mathbf{S}_{\mathbf{W}_1})\mathbf{S}_{\mathbf{W}_1}] - \{\mathbf{f}(\boldsymbol{\nu})\}\boldsymbol{\nu}, \quad \text{and}$$

$$\boldsymbol{\Pi}_2(\mathbf{f}) = 4(E[\mathbf{f}\{E(\mathbf{S}_{\mathbf{W}_1+\mathbf{W}_2}|\mathbf{W}_1)\}E(\mathbf{S}_{\mathbf{W}_1+\mathbf{W}_2}|\mathbf{W}_1)] - \{\mathbf{f}(\boldsymbol{\theta})\}\boldsymbol{\theta}),$$

where $\mathbf{f} \in \mathscr{X}^{**}$. So, $\boldsymbol{\Pi}_1$ and $\boldsymbol{\Pi}_2$ are symmetric positive continuous linear operators. For Banach space valued random elements $\mathbf{U}$ and $\mathbf{V}$ defined on the same probability

space with $\mathbf{U}$ having finite Bochner expectation, the conditional expectation of $\mathbf{U}$ given $\mathbf{V}$ exists and can be properly defined (see Vakhania et al. 1987, pp. 125–128). Let $(\mathbf{X}_i, \mathbf{Y}_i)$, $1 \leq i \leq n$, be i.i.d. paired observations with the $\mathbf{X}_i$'s and the $\mathbf{Y}_i$'s taking values in a smooth Banach space $\mathscr{X}$. The next theorem gives the asymptotic distributions of $\mathbf{T}_S$ and $\mathbf{T}_{SR}$.

Theorem 20.1 *Suppose that the dual space $\mathscr{X}^*$ is a separable and type 2 Banach space. Then, for any probability measure P on $\mathscr{X}$, $n^{1/2}(\mathbf{T}_S - \boldsymbol{\nu})$ converges weakly to a Gaussian limit $G(\mathbf{0}, \boldsymbol{\Pi}_1)$ as $n \to \infty$. Further, if $\mathscr{X}^*$ is p-uniformly smooth for some $p \in (4/3, 2]$, we have weak convergence of $n^{1/2}(\mathbf{T}_{SR} - \boldsymbol{\theta})$ to a Gaussian limit $G(\mathbf{0}, \boldsymbol{\Pi}_2)$ as $n \to \infty$.*

Proof Using the central limit theorem for i.i.d. random elements in a separable and type 2 Banach space (see Araujo and Giné 1980, Theorem 7.5(i)), we get that $n^{1/2}(\mathbf{T}_S - \boldsymbol{\nu})$ converges *weakly* to $G(\mathbf{0}, \boldsymbol{\Pi}_1)$.

Note that $\mathbf{T}_{SR} - \boldsymbol{\theta}$ is a Banach space valued U-statistic with kernel $\mathbf{h}(\mathbf{w}_i, \mathbf{w}_j) = \mathbf{S}_{\mathbf{w}_i + \mathbf{w}_j} - \boldsymbol{\theta}$, which satisfies $E\{\mathbf{h}(\mathbf{W}_i, \mathbf{W}_j)\} = \mathbf{0}$. By the Hoeffding type decomposition for Banach space valued U-statistics (see Borovskikh 1991, p. 430), we have

$$\mathbf{T}_{SR} - \boldsymbol{\theta} = \frac{2}{n} \sum_{i=1}^{n} [E\{\mathbf{S}_{\mathbf{W}_i + \mathbf{W}'} \mid \mathbf{W}_i\} - \boldsymbol{\theta}] + \mathbf{R}_n,$$

where $\mathbf{W}'$ is an independent copy of $\mathbf{W}_1$. So, $\mathbf{R}_n = 2[n(n-1)]^{-1} \sum_{1 \leq i < j \leq n} \tilde{\mathbf{h}}(\mathbf{W}_i, \mathbf{W}_j)$, where $\tilde{\mathbf{h}}(\mathbf{w}_i, \mathbf{w}_j) = \mathbf{h}(\mathbf{w}_i, \mathbf{w}_j) - E\{\mathbf{h}(\mathbf{W}_i, \mathbf{W}_j) \mid \mathbf{W}_i = \mathbf{w}_i\} - E\{\mathbf{h}(\mathbf{W}_i, \mathbf{W}_j) \mid \mathbf{W}_j = \mathbf{w}_j\}$. Note that $||\tilde{\mathbf{h}}(\mathbf{w}_i, \mathbf{w}_j)|| \leq 4$. Using the boundedness of $\tilde{\mathbf{h}}(\cdot, \cdot)$ and Theorem 5.1 in Borovskikh (1991), it follows that for any $q \in [1, p]$,

$$E(||n^{1/2}\mathbf{R}_n||^q) \leq \tilde{\alpha}_q n^{2-(3q/2)} \tag{20.1}$$

for every $n \geq 2$ and a constant $\tilde{\alpha}_q$. Thus, if $p > 4/3$, $E(||n^{1/2}\mathbf{R}_n||^q)$ converges to zero as $n \to \infty$ for any $q \in (4/3, p]$. This implies that $n^{1/2}\mathbf{R}_n$ converges to zero *in probability* as $n \to \infty$.

Now, $n^{-1/2} \sum_{i=1}^{n} E\{\mathbf{h}(\mathbf{W}_i, \mathbf{W}') \mid \mathbf{W}_i\}$ converge *weakly* to $G(\mathbf{0}, \boldsymbol{\Pi}_2)$ as $n \to \infty$ by the central limit theorem for i.i.d. random elements in a separable type 2 Banach space (see Araujo and Giné 1980, Theorem 7.5(i)). This, together with the fact that $n^{1/2}\mathbf{R}_n$ converges to zero *in probability*, completes the proof. □

Let $c_{1\alpha}$ and $c_{2\alpha}$ denote the $(1-\alpha)$ quantiles of the distributions of $||G(\mathbf{0}, \boldsymbol{\Pi}_1)||$ and $||G(\mathbf{0}, \boldsymbol{\Pi}_2)||$, respectively. The test based on $\mathbf{T}_S$ rejects $H_0^{(1)}$ when $||n^{1/2}\mathbf{T}_S|| > c_{1\alpha}$ and the test based on $\mathbf{T}_{SR}$ rejects $H_0^{(2)}$ when $||n^{1/2}\mathbf{T}_{SR}|| > c_{2\alpha}$.

Corollary 20.1 *The asymptotic sizes of these tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ will be the same as their nominal level. Further, these tests are consistent whenever $\boldsymbol{\nu} \neq \mathbf{0}$ and $\boldsymbol{\theta} \neq \mathbf{0}$, respectively. So, if the distribution of $\mathbf{Y}_1 - \mathbf{X}_1$ is symmetric and its*

spatial median is non-zero, then these tests are consistent. In particular, these tests are consistent for location shift alternatives.

We next describe how to compute the critical value of the tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ using their asymptotic distributions under the null hypotheses obtained in Theorem 20.1. Suppose that $\mathscr{X}$ is a separable Hilbert space and $\mathbf{Z}$ is a zero mean Gaussian random element in $\mathscr{X}$ with covariance operator $\mathbf{C}$. Then, it can be shown using the spectral decomposition of the compact self-adjoint operator $\mathbf{C}$ (see Theorem IV.2.4, and Theorem 1.3 and Corollary 2 in pp. 159–160 in Vakhania et al. 1987) that the distribution of $||\mathbf{Z}||^2$ is a weighted sums of independent chi-square variables each with one degree of freedom, where the weights are the eigenvalues of the $\mathbf{C}$. Thus, if $\mathscr{X}$ is a separable Hilbert space, the asymptotic distributions of $||n^{1/2}\mathbf{T}_S||^2$ and $||n^{1/2}\mathbf{T}_{SR}||^2$ are weighted sums of independent chi-square variables each with one degree of freedom, where the weights are the eigenvalues of $\boldsymbol{\Pi}_1$and $\boldsymbol{\Pi}_2$, respectively. The eigenvalues of $\boldsymbol{\Pi}_1$ and $\boldsymbol{\Pi}_2$ can be estimated by the eigenvalues of $\hat{\Pi}_1$ and $\hat{\Pi}_2$, which are defined as

$$\widehat{\boldsymbol{\Pi}}_1 = \frac{1}{n-1}\left\{\sum_{i=1}^{n}\left(\frac{\mathbf{W}_i}{||\mathbf{W}_i||} - \hat{\boldsymbol{\nu}}\right) \otimes \left(\frac{\mathbf{W}_i}{||\mathbf{W}_i||} - \hat{\boldsymbol{\nu}}\right)\right\},$$

$$\widehat{\boldsymbol{\Pi}}_2 = \frac{4}{n-1}\left\{\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{\mathbf{W}_i+\mathbf{W}_j}{||\mathbf{W}_i+\mathbf{W}_j||} - \hat{\boldsymbol{\theta}}\right)\right.$$
$$\left. \otimes \left(\frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{\mathbf{W}_i+\mathbf{W}_j}{||\mathbf{W}_i+\mathbf{W}_j||} - \hat{\boldsymbol{\theta}}\right)\right\}.$$

Here, $\mathbf{x}\otimes\mathbf{x} : \mathscr{X} \to \mathscr{X}$ is the tensor product in the Hilbert space $\mathscr{X}$, which is defined as $\langle(\mathbf{x}\otimes\mathbf{x})(\mathbf{f}), \mathbf{g}\rangle = \langle\mathbf{x}, \mathbf{f}\rangle\langle\mathbf{x}, \mathbf{g}\rangle$ for $\mathbf{f}, \mathbf{g}, \mathbf{x} \in \mathscr{X}$. Further, $\hat{\boldsymbol{\nu}} = n^{-1}\sum_{i=1}^{n}\mathbf{W}_i/||\mathbf{W}_i||$ and $\hat{\boldsymbol{\theta}} = 2\{n(n-1)\}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(\mathbf{W}_i+\mathbf{W}_j)/||\mathbf{W}_i+\mathbf{W}_j||$. The critical values $c_{1\alpha}$ and $c_{2\alpha}$ can be obtained by simulating from the estimated asymptotic distributions of $\mathbf{T}_S$ and $\mathbf{T}_{SR}$. On the other hand, if $\mathscr{X}$ is a general Banach space satisfying the assumptions of Theorem 20.1, we no longer have the weighted chi-square representations for the asymptotic distributions of $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ under the null hypotheses. However, we can estimate $\boldsymbol{\Pi}_1$ and $\boldsymbol{\Pi}_2$ by their empirical counterparts, which are defined in a similar way as the definitions above. We can simulate from the asymptotic Gaussian distributions with the estimated covariance operators to compute the critical values of the tests.

20.3 Asymptotic Distributions Under Shrinking Alternatives

Consider i.i.d. paired observations $(\mathbf{X}_i, \mathbf{Y}_i)$, $1 \le i \le n$, where the $\mathbf{X}_i$'s and the $\mathbf{Y}_i$'s take values in a smooth Banach space $\mathscr{X}$. In this section, we shall derive the asymptotic distribution of $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ under sequences of shrinking location shift alternatives, where $\mathbf{W}_i = \mathbf{Y}_i - \mathbf{X}_i$ is symmetrically distributed about $n^{-1/2}\boldsymbol{\eta}$ for some fixed nonzero $\boldsymbol{\eta} \in \mathscr{X}$ and $1 \le i \le n$. For some of the finite dimensional multivariate extensions of the sign and the signed-rank tests, such alternative hypotheses have been shown to be contiguous to the null and yields nondegenerate asymptotic distributions of the test statistics (see, e.g., Randles 1989; Möttönen et al. 1997; Oja 1999).

Definition 20.8 The norm in $\mathscr{X}$ is said to be twice Gâteaux differentiable at $\mathbf{x} \neq \mathbf{0}$ with Hessian (or second order Gâteaux derivative) $\mathbf{H}_{\mathbf{x}}$, which is a continuous linear map from $\mathscr{X}$ to $\mathscr{X}^*$, if $\lim_{t\to 0}\, t^{-1}(\mathbf{S}_{\mathbf{x}+t\mathbf{h}} - \mathbf{S}_{\mathbf{x}}) = \mathbf{H}_{\mathbf{x}}(\mathbf{h})$ for all $\mathbf{h} \in \mathscr{X}$. Here, the limit is assumed to exist in the norm topology of $\mathscr{X}^*$.

Norms in Hilbert spaces and L_p spaces for $p \in [2, \infty)$ are twice Gâteaux differentiable. We refer to Chapters 4 and 5 in Borwein and Vanderwerff (2010) for further details. For the next theorem, let us assume that the norm in $\mathscr{X}$ is twice Gâteaux differentiable at every $\mathbf{x} \neq \mathbf{0}$, and denote the Hessians of the functions $\mathbf{x} \mapsto E\{||\mathbf{x} - \mathbf{W}_1|| - ||\mathbf{W}_1||\}$ and $\mathbf{x} \mapsto E\{||2\mathbf{x} - \mathbf{W}_1 - \mathbf{W}_2|| - ||\mathbf{W}_1 + \mathbf{W}_2||\}$ by $\mathbf{J}_{\mathbf{x}}^{(1)}$ and $\mathbf{J}_{\mathbf{x}}^{(2)}$, respectively. The following theorem gives the asymptotic distributions of $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ under the sequence of shrinking alternatives mentioned earlier.

Theorem 20.2 *Suppose that $\mathscr{X}^*$ is a separable and type 2 Banach space. Assume that the distribution of $\mathbf{W}_1$ is nonatomic, and $\mathbf{J}_{\mathbf{0}}^{(1)}$ exists. Then, $n^{1/2}\mathbf{T}_S$ converges weakly to a Gaussian limit $G\{\mathbf{J}_{\mathbf{0}}^{(1)}(\boldsymbol{\eta}), \boldsymbol{\Pi}_1\}$ as $n \to \infty$. Further, if $\mathscr{X}^*$ is a p-uniformly smooth Banach space for some $p \in (4/3, 2]$ and $\mathbf{J}_{\mathbf{0}}^{(2)}$ exists, we have weak convergence of $n^{1/2}\mathbf{T}_{SR}$ to a Gaussian limit $G\{\mathbf{J}_{\mathbf{0}}^{(1)}(\boldsymbol{\eta}), \boldsymbol{\Pi}_2\}$. Here, the expectations in the definitions of $\mathbf{J}_{\mathbf{0}}^{(1)}$ and $\mathbf{J}_{\mathbf{0}}^{(2)}$ are taken with respect to the symmetric distribution of $\mathbf{W}_1$ about zero under the null hypothesis.*

Proof We first derive the asymptotic distribution of $\mathbf{T}_{SR}$. Let $\boldsymbol{\eta}_n = n^{-1/2}\boldsymbol{\eta}$. Applying the Hoeffding type decomposition for Banach space valued U-statistics as in the proof of Theorem 20.1, it follows that

$$\mathbf{T}_{SR} - \boldsymbol{\theta}(\boldsymbol{\eta}_n) = \frac{2}{n}\sum_{i=1}^{n}\{E(\mathbf{S}_{\mathbf{W}_i+\mathbf{W}'} \mid \mathbf{W}_i) - \boldsymbol{\theta}(\boldsymbol{\eta}_n)\} + \tilde{\mathbf{R}}_n, \tag{20.2}$$

where $\mathbf{W}'$ is an independent copy of $\mathbf{W}$. Arguing as in the proof of Theorem 20.1, it can be shown that $E(||n^{1/2}\tilde{\mathbf{R}}_n||^q)$ satisfies the bound obtained in (20.1) for every $n \ge 2$ and any $q \in (4/3, p]$. Thus, $n^{1/2}\tilde{\mathbf{R}}_n \to 0$ *in probability* as $n \to \infty$ under the sequence of shrinking shifts.

By definition, $\boldsymbol{\theta}(\boldsymbol{\eta}_n) = E(\mathbf{S}_{\mathbf{W}_1^0+\mathbf{W}_2^0+2\boldsymbol{\eta}_n}) = E(\mathbf{S}_{2\boldsymbol{\eta}_n-\mathbf{W}_1^0-\mathbf{W}_2^0})$, where $\mathbf{W}_i^0 = \mathbf{W}_i - \boldsymbol{\eta}_n$ has the same distribution as that of $\mathbf{W}_i$ under the null hypothesis for $1 \le i \le n$. So, $n^{1/2}\boldsymbol{\theta}(\boldsymbol{\eta}_n)$ converges to $\mathbf{J}_{\mathbf{0}}^{(2)}(\boldsymbol{\eta})$ as $n \to \infty$.

Define $\boldsymbol{\Phi}_n(\mathbf{W}_i) = 2n^{-1/2}\{E(\mathbf{S}_{\mathbf{W}_i+\mathbf{W}'} \mid \mathbf{W}_i) - \boldsymbol{\theta}(\boldsymbol{\eta}_n)\}$. So, $E\{\boldsymbol{\Phi}_n(\mathbf{W}_i)\} = \mathbf{0}$. To prove the asymptotic Gaussianity of $\sum_{i=1}^n \boldsymbol{\Phi}_n(\mathbf{W}_i)$, it is enough to show that the triangular array $\{\boldsymbol{\Phi}_n(\mathbf{W}_1), \boldsymbol{\Phi}_n(\mathbf{W}_2), \ldots, \boldsymbol{\Phi}_n(\mathbf{W}_n)\}_{n=1}^{\infty}$ of rowwise i.i.d. random elements satisfy the conditions of Corollary 7.8 in Araujo and Giné (1980).

Observe that for any $\epsilon > 0$,

$$\sum_{i=1}^{n} P\{||\boldsymbol{\Phi}_n(\mathbf{W}_i)|| > \epsilon\} \le \frac{8}{(\epsilon n)^{3/2}} \sum_{i=1}^{n} E\{||E(\mathbf{S}_{\mathbf{W}_i+\mathbf{W}'} \mid \mathbf{W}_i) - \boldsymbol{\theta}(\boldsymbol{\eta}_n)||^3\} \le \frac{64}{(\epsilon^3 n)^{1/2}}.$$

Thus, $\lim_{n\to\infty} \sum_{i=1}^n P\{||\boldsymbol{\Phi}_n(\mathbf{W}_i)|| > \epsilon\} = 0$ for every $\epsilon > 0$, which ensures that condition (1) of Corollary 7.8 in Araujo and Giné (1980) holds.

We next verify condition (2) of Corollary 7.8 in Araujo and Giné (1980). Let us fix $\mathbf{f} \in \mathscr{X}$. Since $||\mathbf{S}_{\mathbf{x}}|| = 1$ for all $\mathbf{x} \ne \mathbf{0}$, we can choose $\delta = 1$ in that condition (2). Then, using the linearity of $\mathbf{f}$, we have

$$\sum_{i=1}^{n} E[\mathbf{f}^2\{\boldsymbol{\Phi}_n(\mathbf{W}_i)\}] = 4n^{-1} \sum_{i=1}^{n} E[\{V_{n,i} - E(V_{n,i})\}^2], \tag{20.3}$$

where $V_{n,i} = \mathbf{f}\{E(\mathbf{S}_{\mathbf{W}_i+\mathbf{W}'} \mid \mathbf{W}_i)\}$. Since the $\mathbf{W}_i$'s are identically distributed, the right-hand side of (20.3) equals $4E[\{V_{n,1} - E(V_{n,1})\}^2]$. Note that $V_{n,1} = \mathbf{f}\{E(\mathbf{S}_{\mathbf{W}_1+\mathbf{W}_2^0+\boldsymbol{\eta}_n} \mid \mathbf{W}_1)\}$. Since the norm in $\mathscr{X}$ is assumed to be twice Gâteaux differentiable, it follows from Theorem 4.6.15(a) and Proposition 4.6.16 in Borwein and Vanderwerff (2010) that the norm in $\mathscr{X}$ is Fréchet differentiable. This in turn implies that the map $\mathbf{x} \mapsto \mathbf{S}_{\mathbf{x}}$ is continuous on $\mathscr{X}\setminus\{\mathbf{0}\}$ (see Borwein and Vanderwerff 2010, Corollary 4.2.12). Using this fact, it follows from the dominated convergence theorem for Banach space valued random elements that

$$\begin{aligned} E(\mathbf{S}_{\mathbf{W}_1+\mathbf{W}_2^0+\boldsymbol{\eta}_n} \mid \mathbf{W}_1 = \mathbf{w}_1) &= E(\mathbf{S}_{\mathbf{w}_1+\mathbf{W}_2^0+\boldsymbol{\eta}_n}) \\ &\longrightarrow E(\mathbf{S}_{\mathbf{w}_1+\mathbf{W}_2^0}) = E(\mathbf{S}_{\mathbf{W}_1^0+\mathbf{W}_2^0} \mid \mathbf{W}_1^0 = \mathbf{w}_1) \end{aligned} \tag{20.4}$$

as $n \to \infty$ for almost all values of $\mathbf{w}_1$. Thus, $E(V_{n,1})$ converges to $E[\mathbf{f}\{E(\mathbf{S}_{\mathbf{W}_1^0+\mathbf{W}_2^0} \mid \mathbf{W}_1^0)\}]$ as $n \to \infty$ by the usual dominated convergence theorem. Similarly, $E(V_{n,1}^2)$ converges to $E[\mathbf{f}^2\{E(\mathbf{S}_{\mathbf{W}_1^0+\mathbf{W}_2^0} \mid \mathbf{W}_1^0)\}]$ as $n \to \infty$. So, $\sum_{i=1}^n E[\mathbf{f}^2\{\boldsymbol{\Phi}_n(\mathbf{W}_i)\}] \to \boldsymbol{\Pi}_2(\mathbf{f}, \mathbf{f})$ as $n \to \infty$, where $\boldsymbol{\Pi}_2$ is as defined before Theorem 20.1 in Sect. 20.2 and the expectations in that definition are taken under the null hypothesis. This completes the verification of condition (2) of Corollary 7.8 in Araujo and Giné (1980).

Finally, for the verification of condition (3) of Corollary 7.8 in Araujo and Giné (1980), suppose that $\{\mathscr{F}_k\}_{k\geq 1}$ is a sequence of finite dimensional subspaces of $\mathscr{X}^*$ such that $\mathscr{F}_k \subseteq \mathscr{F}_{k+1}$ for all $k \geq 1$, and the closure of $\bigcup_{k=1}^{\infty} \mathscr{F}_k$ is $\mathscr{X}^*$. Such a sequence of subspaces exists because of the separability of $\mathscr{X}^*$. For any $\mathbf{x} \in \mathscr{X}^*$ and any $k \geq 1$, we define $d(\mathbf{x}, \mathscr{F}_k) = \inf\{||\mathbf{x} - \mathbf{y}|| : \mathbf{y} \in \mathscr{F}_k\}$. It is straightforward to verify that for every $k \geq 1$, the map $\mathbf{x} \mapsto d(\mathbf{x}, \mathscr{F}_k)$ is continuous and bounded on any closed ball in $\mathscr{X}^*$. Thus, using (20.4), it follows that $\boldsymbol{\theta}(\boldsymbol{\eta}_n) \to 0$ as $n \to \infty$, and

$$\begin{aligned}\sum_{i=1}^{n} E[d^2\{\boldsymbol{\Phi}_n(\mathbf{W}_i), \mathscr{F}_k\}] &= 4n^{-1}\sum_{i=1}^{n} E[d^2\{E(\mathbf{S}_{\mathbf{W}_i+\mathbf{W}'+\boldsymbol{\eta}_n} \mid \mathbf{W}_i) - \boldsymbol{\theta}(\boldsymbol{\eta}_n), \mathscr{F}_k\}] \\ &= 4E[d^2\{E(\mathbf{S}_{\mathbf{W}_1+\mathbf{W}_2^0+\boldsymbol{\eta}_n} \mid \mathbf{W}_1) - \boldsymbol{\theta}(\boldsymbol{\eta}_n), \mathscr{F}_k\}] \\ &\longrightarrow 4E[d^2\{E(\mathbf{S}_{\mathbf{W}_1^0+\mathbf{W}_2^0} \mid \mathbf{W}_1^0), \mathscr{F}_k\}]\end{aligned}$$

as $n \to \infty$. From the choice of the $\mathscr{F}_k$'s, it can be shown that $d(\mathbf{x}, \mathscr{F}_k) \to 0$ as $k \to \infty$ for all $\mathbf{x} \in \mathscr{X}^*$. So, we have $\lim_{k\to\infty} E[d^2\{E(\mathbf{S}_{\mathbf{W}_1^0+\mathbf{W}_2^0} \mid \mathbf{W}_1^0), \mathscr{F}_k\}] = 0$, and this completes the verification of condition (3) of Corollary 7.8 in Araujo and Giné (1980).

Thus, $\sum_{i=1}^{n} \boldsymbol{\Phi}_n(\mathbf{W}_i)$ converges *weakly* to a zero mean Gaussian distribution in $\mathscr{X}$ as $n \to \infty$. Further, its asymptotic covariance is $\boldsymbol{\Pi}_2$, which was obtained while checking condition (2) of Corollary 7.8 in Araujo and Giné (1980). Thus, it follows from equation (20.2) at the beginning of the proof that

$$n^{1/2}\{\mathbf{T}_{SR} - \boldsymbol{\theta}(\boldsymbol{\eta}_n)\} \longrightarrow G(\mathbf{0}, \boldsymbol{\Pi}_2)$$

weakly as $n \to \infty$ under the sequence of shrinking shifts. The above weak convergence and the fact that $n^{1/2}\boldsymbol{\theta}(\boldsymbol{\eta}_n)$ converges to $\mathbf{J}_{\mathbf{0}}^{(2)}(\boldsymbol{\eta})$ as $n \to \infty$ complete the derivation of the asymptotic distribution of $\mathbf{T}_{SR}$.

Since $\mathbf{T}_S$ is a sum of independent random elements, its asymptotic distribution can be obtained by using arguments similar to those used to derive the asymptotic distribution of $\sum_{i=1}^{n} \boldsymbol{\Phi}_n(\mathbf{W}_i)$ given above. It follows that $n^{1/2}\mathbf{T}_S$ has an asymptotic Gaussian distribution with mean $\mathbf{J}_{\mathbf{0}}^{(1)}(\boldsymbol{\eta})$ and covariance $\boldsymbol{\Pi}_1$ as $n \to \infty$ under the sequence of shrinking alternatives considered in the theorem. □

Let $\mathscr{X}$ be a separable Hilbert space and $\mathbf{Y}_1 - \mathbf{X}_1 = \sum_{k=1}^{\infty} Z_k \boldsymbol{\phi}_k$ for an orthonormal basis $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots$ of $\mathscr{X}$ and the Z_k's are real-valued random variables. Then, the expectations defining $\mathbf{J}_{\mathbf{0}}^{(1)}$ and $\mathbf{J}_{\mathbf{0}}^{(2)}$ are finite if any two dimensional marginal of $(Z_1, Z_2, \ldots)$ has a density that is bounded on bounded subsets of $\mathbb{R}^2$. If $\mathscr{X}$ is an L_p space for some $p \in (1, \infty)$, then $\mathbf{J}_{\mathbf{0}}^{(1)}$ and $\mathbf{J}_{\mathbf{0}}^{(2)}$ exist if $E(||\mathbf{W}_1||^{-1})$ and $E(||\mathbf{W}_1 + \mathbf{W}_2||^{-1})$ are finite.

We next compare the asymptotic powers of the tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ with some paired sample mean based tests for functional data in $L_2[a, b]$, where $a, b \in \mathbb{R}$. Cuevas et al. (2004) studied a test for analysis of variance in $L_2[a, b]$, and the test

statistic for the two sample problem is based on $||\overline{\mathbf{X}} - \overline{\mathbf{Y}}||^2$. We consider the natural paired sample analog of this statistic and define it as $T_1 = ||\overline{\mathbf{W}}||^2$. Horváth et al. (2013) studied a couple of two sample tests in $L_2[a, b]$ based on the projections of the sample functions onto the subspace formed by finitely many eigenfunctions of the sample pooled covariance operator. We consider the paired sample analogs of these tests, and the corresponding test statistics are defined as $T_2 = \sum_{k=1}^{L} (\langle \overline{\mathbf{W}}, \hat{\boldsymbol{\psi}}_k \rangle)^2$ and $T_3 = \sum_{k=1}^{L} \hat{\lambda}_k^{-1} (\langle \overline{\mathbf{W}}, \hat{\boldsymbol{\psi}}_k \rangle)^2$. Here, the $\hat{\lambda}_k$'s denote the eigenvalues of the empirical covariance of the $\mathbf{W}_i$'s in descending order of magnitudes, and the $\hat{\boldsymbol{\psi}}_k$'s are the corresponding empirical eigenfunctions. Consider i.i.d. paired observations $(\mathbf{X}_i, \mathbf{Y}_i)$, $1 \le i \le n$, where the $\mathbf{X}_i$'s and the $\mathbf{Y}_i$'s take values in $L_2[a, b]$. The next theorem gives the asymptotic distributions of T_1, T_2, and T_3 under sequences of shrinking location shift alternatives, where $\mathbf{W}_i = \mathbf{Y}_i - \mathbf{X}_i$ is symmetrically distributed about $n^{-1/2}\boldsymbol{\eta}$ for some fixed nonzero $\boldsymbol{\eta} \in \mathscr{X}$ and $1 \le i \le n$.

Theorem 20.3

(a) If $E(||\mathbf{W}_1||^2) < \infty$, nT_1 converges weakly to $\sum_{k=1}^{\infty} \lambda_k \chi^2_{(1)}(\beta_k^2/\lambda_k)$ as $n \to \infty$. Here, the λ_k's are the eigenvalues of the covariance operator $\boldsymbol{\Sigma}$ of $\mathbf{W}_1$ in decreasing order of magnitudes, the $\boldsymbol{\psi}_k$'s are the eigenfunctions corresponding to the λ_k's, $\beta_k = \langle \boldsymbol{\eta}, \boldsymbol{\psi}_k \rangle$, and $\chi^2_{(1)}(\beta_k^2/\lambda_k)$ denotes the non-central chi-square variable with one degree of freedom and non-centrality parameter β_k^2/λ_k for $k \ge 1$.

(b) Suppose that for some $L \ge 1$, we have $\lambda_1 > \ldots > \lambda_L > \lambda_{L+1} > 0$. Assume that $E(||\mathbf{W}_1||^4) < \infty$. Then, nT_2 converges weakly to $\sum_{k=1}^{L} \lambda_k \chi^2_{(1)}(\beta_k^2/\lambda_k)$, and nT_3 converges weakly to $\sum_{k=1}^{L} \chi^2_{(1)}(\beta_k^2/\lambda_k)$ as $n \to \infty$.

Proof

(a) As in the proof of Theorem 20.2, we denote $\boldsymbol{\eta}_n = n^{-1/2}\boldsymbol{\eta}$, and $\mathbf{W}_i^0 = \mathbf{W}_i - \boldsymbol{\eta}_n$ has the same distribution as that of $\mathbf{W}_i$ under the null hypothesis for $1 \le i \le n$. Now, by the central limit theorem for i.i.d. random elements in a separable Hilbert space (see Araujo and Giné 1980, Theorem 7.5(i)), it follows that $n^{1/2}\overline{\mathbf{W}^0}$ converges *weakly* to $G(\mathbf{0}, \boldsymbol{\Sigma})$ as $n \to \infty$. Thus, $n^{1/2}\overline{\mathbf{W}} = n^{1/2}(\overline{\mathbf{W}^0} + \boldsymbol{\eta}_n)$, converges *weakly* to $G(\boldsymbol{\eta}, \boldsymbol{\Sigma})$ as $n \to \infty$. Now, the distribution of $||G(\boldsymbol{\eta}, \boldsymbol{\Sigma})||^2$ is the same as that of $\sum_{k=1}^{\infty} \lambda_k \chi^2_{(1)}(\beta_k^2/\lambda_k)$ using the spectral decomposition of the compact self-adjoint operator $\boldsymbol{\Sigma}$. This proves part (a) of the proposition.

(b) Let $\mathbf{V} = (\langle \overline{\mathbf{W}}, \boldsymbol{\psi}_1 \rangle, \langle \overline{\mathbf{W}}, \boldsymbol{\psi}_2 \rangle, \ldots, \langle \overline{\mathbf{W}}, \boldsymbol{\psi}_L \rangle)^\top$ and $\tilde{\boldsymbol{\beta}} = (\beta_1, \ldots, \beta_L)^\top$. It follows from the central limit theorem in $\mathbb{R}^L$ that the distribution of $n^{1/2}(\mathbf{V} - n^{-1/2}\tilde{\boldsymbol{\beta}})$ converges *weakly* to the L-variate Gaussian distribution $N_L(\mathbf{0}, \boldsymbol{\Lambda}_L)$ as $n \to \infty$ under the sequence of shrinking shifts considered, where $\boldsymbol{\Lambda}_L$ is the diagonal matrix $Diag(\lambda_1, \ldots, \lambda_L)$. Thus, under the given sequence of shifts, the distribution of $n^{1/2}\mathbf{V}$ converges *weakly* to the L-variate Gaussian distribution $N_L(\tilde{\boldsymbol{\beta}}, \boldsymbol{\Lambda}_L)$ distribution as $n \to \infty$.

From arguments similar to those in the proof of Theorem 5.3 in Horváth and Kokoszka (2012), and using the assumptions in the present theorem, we get

$$\max_{1\le k\le L} n^{1/2}|\langle \overline{\mathbf{W}}, \hat{\boldsymbol{\psi}}_k - \hat{c}_k \boldsymbol{\psi}_k\rangle| = o_P(1) \tag{20.5}$$

as $n \to \infty$ under the sequence of shrinking shifts. Here $\hat{\boldsymbol{\psi}}_k$ is the empirical version of $\boldsymbol{\psi}_k$ and $\hat{c}_k = sign(\langle \hat{\boldsymbol{\psi}}_k, \boldsymbol{\psi}_k\rangle)$. In view of (20.5), the limiting distribution of $n\sum_{k=1}^{L}(\langle \overline{\mathbf{W}}, \hat{\boldsymbol{\psi}}_k\rangle)^2$ is the same as that of $n\sum_{k=1}^{L}(\langle \overline{\mathbf{W}}, \hat{c}_k\boldsymbol{\psi}_k\rangle)^2 = n||\mathbf{V}||^2$, and the latter converges *weakly* to $||N_L(\tilde{\boldsymbol{\beta}}, \boldsymbol{\Lambda}_L)||^2$ as $n \to \infty$. Thus, nT_2 converges *weakly* to $\sum_{k=1}^{L}\lambda_k\chi^2_{(1)}(\beta_k^2/\lambda_k)$ as $n \to \infty$ under the sequence of shrinking shifts considered.

It also follows using similar arguments as in the proof of Theorem 5.3 in Horváth and Kokoszka (2012) that under the assumptions of the present theorem, we have

$$\max_{1\le k\le L} n^{1/2}\hat{\lambda}_k^{-1/2}|\langle \overline{\mathbf{W}}, \hat{\boldsymbol{\psi}}_k - \hat{c}_k \boldsymbol{\psi}_k\rangle| = o_P(1)$$

as $n \to \infty$ under the given sequence of shrinking shifts. Similar arguments as in the case of T_2 yield the asymptotic distribution of nT_3, and this completes the proof. □

20.3.1 Comparison of Asymptotic Powers of Different Tests

We now compare the asymptotic powers of the tests based on $\mathbf{T}_S$, $\mathbf{T}_{SR}$, T_1, T_2, and T_3 under shrinking location shifts. For this we have taken the random element $\mathbf{W}_i \in L_2[0,1]$ to have the same distribution as that of $\sum_{k=1}^{\infty} 2^{1/2}\{(k-0.5)\pi\}^{-1}Z_k sin\{(k-0.5)\pi t\}$, where the Z_k's are independent random variables for $k \ge 1$. We have considered two distributions of the Z_k's, namely, the Z_k's having $N(0,1)$ distributions, and $Z_k = U_k(V/5)^{-1/2}$ with the U_k's having $N(0,1)$ distributions and V having a chi-square distribution with five degrees of freedom independent of the U_k's for each $k \ge 1$. The latter choice is made to compare the performance of the tests based on $\mathbf{T}_{SR}$ and $\mathbf{T}_S$ with the mean based tests that use T_1, T_2 and T_3, when the underlying distribution is heavy tailed. The two distributions of $\mathbf{W}_i$ considered here correspond to the Karhunen-Loève expansions of the standard Brownian motion and the t process (see Yu et al. 2007) on $[0,1]$ with five degrees of freedom having zero mean and covariance kernel $K(t,s) = \min(t,s)$, $t,s \in [0,1]$, respectively. We call them the sBm and the t(5) distributions, respectively. We have chosen five degrees of freedom for the t distribution so that the finiteness of the fourth moment required in Theorem 20.3 is satisfied. In a finite sample study in the next section, we will also consider t processes with one and three degrees of freedom, which violate some of the moment assumptions in Theorem 20.3. We have taken three choices of the location shift $\boldsymbol{\eta}$, namely, $\boldsymbol{\eta}_1(t) = c$, $\boldsymbol{\eta}_2(t) = ct$ and $\boldsymbol{\eta}_3(t) = ct(1-t)$, where $t \in [0,1]$ and $c > 0$. The plots of the shifts are given in Fig. 20.1 below.

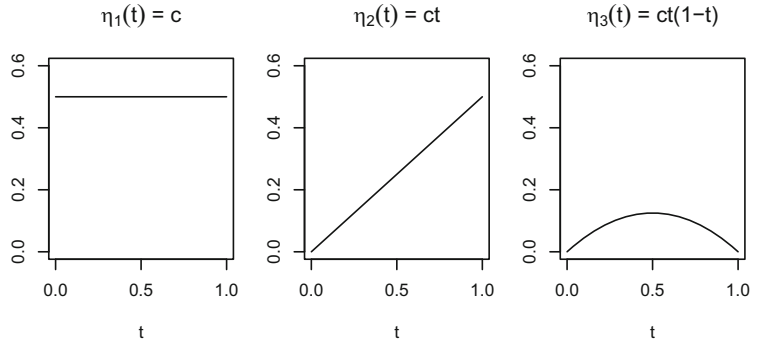

Fig. 20.1 Plots of the shift alternatives

For evaluating the asymptotic powers of these tests, we have used Theorems 20.2 and 20.3. For each of the two distributions of $\mathbf{W}_i$, we have generated 5000 sample functions from it. The operators $\boldsymbol{\Pi}_1$ and $\boldsymbol{\Pi}_2$ are estimated as described in Sect. 20.2 using this sample, and the operator $\boldsymbol{\Sigma}$ is estimated by the sample covariance operator. The eigenvalues and the eigenfunctions of $\boldsymbol{\Pi}_1$, $\boldsymbol{\Pi}_2$, and $\boldsymbol{\Sigma}$ are then estimated by the eigenvalues and the eigenfunctions of the estimates of these operators. We have estimated $\mathbf{J}_{\mathbf{0}}^{(1)}(\boldsymbol{\eta})$ and $\mathbf{J}_{\mathbf{0}}^{(2)}(\boldsymbol{\eta})$ by their sample analogs. The asymptotic powers of the tests are then computed through 1000 Monte Carlo simulations from the asymptotic Gaussian distributions with the estimated parameters. We have used the cumulative variance method described in Horváth et al. (2013) to compute the number L associated with the tests based on T_2 and T_3.

We now discuss the asymptotic powers of different tests under the sBm distribution. It is seen from Fig. 20.2 that the tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ asymptotically outperform the tests based on T_1 and T_2 for the shifts $\boldsymbol{\eta}_1(t)$ and $\boldsymbol{\eta}_3(t)$. The test based on $\mathbf{T}_{SR}$ is asymptotically more powerful than the test based on T_3 for all large values of c in the case of the shift $\boldsymbol{\eta}_1(t)$. However, for the shift $\boldsymbol{\eta}_3(t)$, the test based on T_3 is asymptotically more powerful than the test based on $\mathbf{T}_{SR}$. For the shift $\boldsymbol{\eta}_2(t)$ under the sBm distribution all the tests considered except the test based on $\mathbf{T}_S$ have similar asymptotic powers, and the latter test is asymptotically less powerful than the other competitors under the sBm distribution.

We next consider the t(5) distribution. The tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ asymptotically outperform all the competing tests for all the models considered, except the test based on T_3 for the shift $\boldsymbol{\eta}_3(t)$. The heavy tails of the t(5) distribution adversely effect the performance of the mean based tests, but the tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$, which use spatial signs and spatial signed ranks, are less affected. For the shift $\boldsymbol{\eta}_3(t)$ under the t(5) distribution, although the test based on T_3 asymptotically outperforms the tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$, its performance degrades significantly in comparison to its performance under the sBm distribution, which has lighter tails.

Between the two tests proposed in this paper, the test based on $\mathbf{T}_{SR}$ is asymptotically more powerful than the test based on $\mathbf{T}_S$ for both the distributions and the three

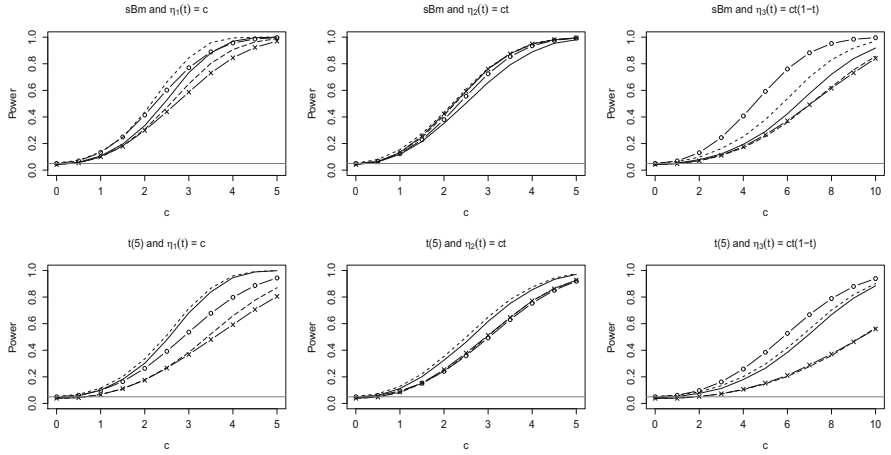

Fig. 20.2 Plots of the asymptotic powers of the tests based on $\mathbf{T}_S$ (—), $\mathbf{T}_{SR}$ (- - -), T_1 (– – –), T_2 (– × –) and T_3 (– ∘ –) for the sBm and the t(5) distributions under shrinking location shifts

shift alternatives considered in this paper. However, the powers of the two tests are quite close under the t(5) distribution.

20.4 Comparison of Finite Sample Powers of Different Tests

We now carry out a comparative study of the finite sample empirical powers of the tests considered in the previous section for location shift alternatives. We consider the distribution of $\mathbf{W}_i$ as in Sect. 20.3, i.e., $\mathbf{W}_i$ has the same distribution as that of $\sum_{k=1}^{\infty} 2^{1/2}\{(k-0.5)\pi\}^{-1} Z_k sin\{(k-0.5)\pi t\}$, where the Z_k's are independent random variables for $k \geq 1$. We have considered four distributions for the $\mathbf{W}_i$'s, namely, the sBm and the t(5) distributions used in Sect. 20.3 as well as the t(1) and the t(3) distributions. For the t(1) (respectively, t(3)) distribution, $Z_k = U_k/(V/r)^{1/2}$, where the U_k's are independent $N(0, 1)$ variables and V has a chi-square distribution with $r = 1$ (respectively, $r = 3$) degree of freedom independent of the U_k's for each $k \geq 1$. The t(1) and the t(3) distributions are chosen to investigate the performance of the mean based tests when the moment conditions required by them fail to hold. We have chosen $n = 20$, and each sample curve is observed at 250 equispaced points in $[0, 1]$. We consider the same three location shifts that were considered in Sect. 20.3.1, namely, $\boldsymbol{\eta}_1(t) = c$, $\boldsymbol{\eta}_2(t) = ct$ and $\boldsymbol{\eta}_3(t) = ct(1-t)$ for $t \in [0, 1]$ and $c > 0$. All the sizes and the powers are evaluated by averaging over 1000 Monte-Carlo simulations. The estimated standard errors of the empirical sizes (respectively, powers) of different tests are of the order of 10^{-3} (respectively, 10^{-2} or less) for all the distributions considered. The empirical power curves of the tests are plotted in Figs. 20.3 and 20.4.

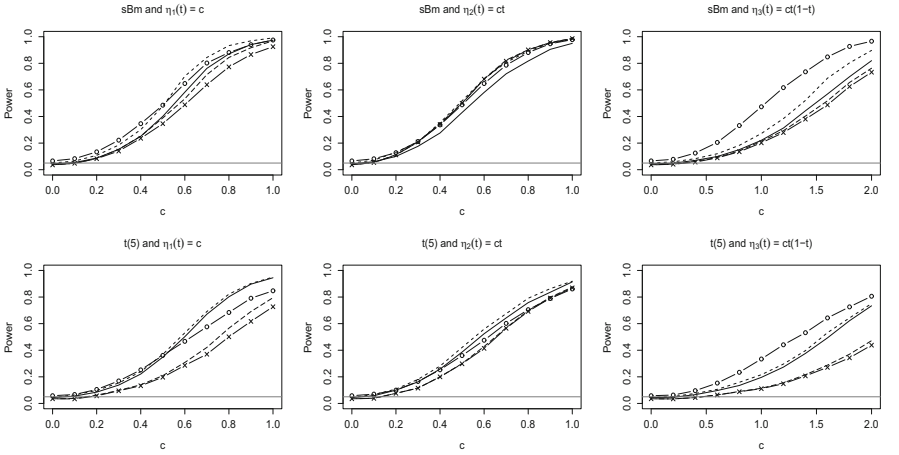

Fig. 20.3 Plots of the empirical powers of the tests based on $\mathbf{T}_S$ (—), $\mathbf{T}_{SR}$ (- - -), T_1 (– – –), T_2 (– × –) and T_3 (– ∘ –) for the sBm and the t(5) distributions

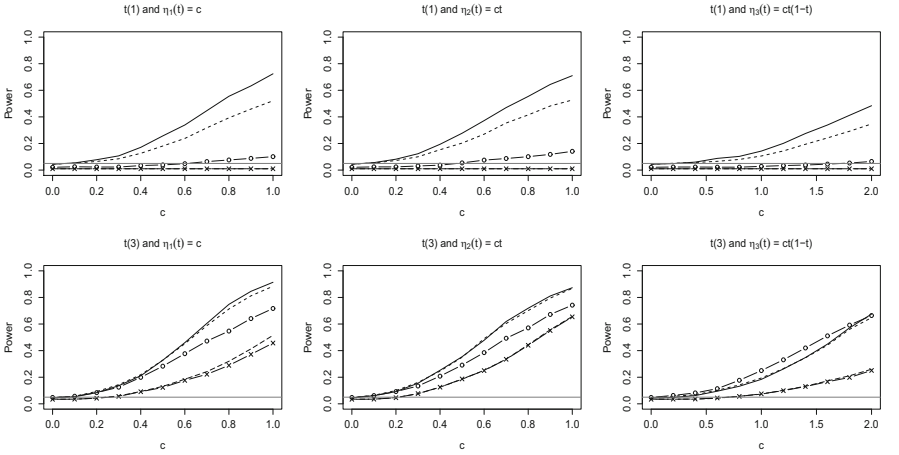

Fig. 20.4 Plots of the asymptotic powers of the tests based on $\mathbf{T}_S$ (—), $\mathbf{T}_{SR}$ (- - -), T_1 (– – –), T_2 (– × –) and T_3 (– ∘ –) for the t(1) and the t(3) distributions

For each of the tests, the difference between its observed size and the nominal 5 % level is statistically insignificant under the sBm, the t(3) and the t(5) distributions. However, the sizes of the tests based on T_1, T_2, and T_3 are 1 %, 1 %, and 2.1 % for the t(1) distribution, all of which are significantly lower than the nominal level. On the other hand, the sizes of the tests based on spatial signs and signed ranks are not significantly different from the nominal level even under t(1) distribution.

We first discuss the performance of different tests under the sBm distribution. The test based on $\mathbf{T}_{SR}$ is significantly more powerful than the tests based on T_1 and

T_2 for the shifts $\boldsymbol{\eta}_1(t)$ and $\boldsymbol{\eta}_3(t)$. The power of the test based on $\mathbf{T}_S$ is not significantly different from that of the test based on T_1 for these two shifts although the former test has slightly more power than the latter test in our simulations. The test based on $\mathbf{T}_S$ significantly outperforms the test based on T_2 for these two shifts. The test based on $\mathbf{T}_{SR}$ is significantly more powerful than the test based on T_3 for large values of c in $\eta_1(t)$. The test based on T_3 significantly outperforms the tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ for the shift $\boldsymbol{\eta}_3(t)$. The power of the test based on $\mathbf{T}_{SR}$ is not significantly different from the powers of the tests based on T_1, T_2, and T_3 for the shift $\boldsymbol{\eta}_2(t)$, but the test based on $\mathbf{T}_S$ is significantly less powerful than all its competitors for this shift under the sBm distribution.

We next consider the t(5) distribution. The tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ are significantly more powerful than the tests based on T_1 and T_2 for all the three shift alternatives considered. The tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ are significantly more powerful than the test based on T_3 for large values of c in the shifts $\boldsymbol{\eta}_1(t)$ and $\boldsymbol{\eta}_2(t)$. Like in the case of the sBm distribution, the test based on T_3 is significantly more powerful than the tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ for the shift $\eta_3(t)$. As in the asymptotic power study in Sect. 20.3, the performance of the mean based tests degrades significantly under the heavy tailed t(5) distribution, while the tests based on spatial signs and signed ranks are less affected. The findings of this finite sample power study under the sBm and the t(5) distributions are very similar to those of the asymptotic power study.

The tests based on T_1, T_2, and T_3 have very low powers for all the shift alternatives considered under the t(1) distribution and are significantly outperformed by the tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$. The non-existence of moments of the t(1) distribution severely affects the performance of the mean based tests, but the tests based on spatial signs and signed ranks are relatively less affected.

Under the t(3) distribution, the tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ significantly outperform the other competitors for the shifts $\boldsymbol{\eta}_1(t)$ and $\boldsymbol{\eta}_2(t)$. The test based on T_3 has significantly more power than the tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ only for $c \in (1, 1.5)$ in the shift $\boldsymbol{\eta}_3(t)$. However, the two latter tests significantly outperform the tests based on T_1 and T_2 for the shift $\boldsymbol{\eta}_3(t)$.

Among the two proposed tests, for all the shifts considered, the spatial sign test based on $\mathbf{T}_S$ is significantly more powerful than the spatial signed-rank test based on $\mathbf{T}_{SR}$ for the very heavy tailed t(1) distribution, while it is significantly less powerful under the light tailed sBm distribution. The powers of these two tests are not significantly different for the t(3) and the t(5) distributions. However, for the t(3) distribution, the spatial sign test is slightly more powerful than the spatial signed-rank test for all the shifts considered, while the situation is reversed for the t(5) distribution. These observations are similar to the relative performance of the spatial sign and signed-rank tests for finite dimensional data under multivariate t distributions (see Möttönen et al. 1997).

20.4.1 Robustness Study of Different Tests

It is known that the univariate sign and signed-rank tests are robust against the presence of outliers in the data, unlike some of the mean based tests like the t test. The influence functions of the sign and signed-rank tests are bounded, and as a result, the sizes and the powers of these tests, even under moderately high contamination proportions, are not much different from their powers in the uncontaminated case (see Hampel et al. 1986, Chapter 3). The definition of the influence function for a test as discussed in Hampel et al. (1986, Chapter 3, p. 191) can be extended to the infinite dimensional setup by using the notion of a Gâteaux derivative. It can be shown that the influence functions of the tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ are given by $[\mathbf{J}_{\mathbf{0}}^{(1)}]^{-1}(\mathbf{S}_{\mathbf{w}})$ and $2[\mathbf{J}_{\mathbf{0}}^{(2)}]^{-1}\{E(\mathbf{S}_{\mathbf{w}+\mathbf{w}_1})\}$ provided the inverses exist, where $\mathbf{w} \in \mathscr{X}$, and $\mathbf{J}_{\mathbf{0}}^{(1)}$ and $\mathbf{J}_{\mathbf{0}}^{(2)}$ are defined before Theorem 20.2. When $\mathscr{X}$ is a separable Hilbert space, both of these influence functions are bounded in norm under the assumptions analogous to those given in Proposition 2.1 in Cardot et al. (2013). So, it is expected that the tests based on $\mathbf{T}_S$ and $\mathbf{T}_{SR}$ will be robust in such cases.

We now conduct an empirical study to assess the robustness of spatial sign and signed-rank tests proposed in this paper in comparison with the mean based tests using T_1, T_2, and T_3. Let the distribution of $\mathbf{Y}_1 - \mathbf{X}_1 \in L_2[0,1]$ be of the form $(1-\epsilon)P + \epsilon Q$, where P is the sBm distribution considered earlier, Q is the Brownian motion with covariance kernel $K(t,s) = 16\min(t,s)$, $t,s \in [0,1]$, and assume that the contamination proportion ϵ takes values $1/20$, $3/20$ and $5/20$. So, even under contamination, *the null hypothesis remains unchanged*. As before, the sample size is chosen to be $n = 20$, and each sample curve is observed at 250 equispaced points in $[0,1]$. For computing the powers of the tests in the presence of outliers in the data, we have considered the location shift $\boldsymbol{\Delta}(t) = 0.8t$, $t \in [0,1]$. This choice ensures that the powers of the tests, even under contamination, are not too close to the nominal 5 % level nor too close to one, and thus a meaningful comparison between the tests can be made. The following table gives the sizes and the powers of different tests, which are evaluated by averaging over 1000 Monte-Carlo simulations, for various contamination models considered.

It is seen from Table 20.1 that except the tests based on T_1 and T_2, the sizes of the other tests considered are close to the nominal 5 % level in the contaminated as well as the uncontaminated situations. The sizes of the mean based tests that

Table 20.1 Sizes and powers of some tests at nominal 5 % level

	Size					Power		
ϵ	$\mathbf{T}_S$	$\mathbf{T}_{SR}$	T_1	T_2	T_3	$\mathbf{T}_S$	$\mathbf{T}_{SR}$	T_3
0	0.044	0.05	0.037	0.036	0.066	0.817	0.891	0.881
1/20	0.053	0.056	0.118	0.115	0.057	0.777	0.818	0.722
3/20	0.044	0.046	0.288	0.281	0.04	0.665	0.616	0.446
5/20	0.044	0.05	0.397	0.37	0.043	0.521	0.423	0.273

use T_1 and T_2 are much larger than the nominal level under contamination though the null hypothesis remains valid. This is probably because the asymptotic critical values of these two tests are under-estimated in the presence of contamination in the data. Interestingly, the size of the mean based test using T_3 is unaffected under contamination. It seems that the standardization involved in the statistic T_3 keeps the size of the test under control even when there are outliers in the data.

For our power study in the presence of outliers in the data, we have excluded the tests based on T_1 and T_2 since they have very high sizes under contamination. The powers of all the tests decrease from their powers in the uncontaminated situation since the contamination increases the variability in the sample. The powers of the proposed spatial sign and signed-rank tests are significantly higher than the power of the mean based test using T_3 for all the contamination models considered. Recall that the test based on T_3 was significantly more powerful than the test based on $\mathbf{T}_S$, and its power was not significantly different from that of the test based on $\mathbf{T}_{SR}$ for the same location shift in the uncontaminated case.

The powers of the tests based on spatial sign and signed rank are comparable when the contamination proportion is at most $3/20$. However, when the proportion of contamination is $5/20$, the spatial sign test becomes significantly more powerful than the spatial signed-rank test. This behavior of the spatial sign and signed-rank tests is similar to that of the univariate sign and signed-rank tests, where it is known that the sign test is more robust than the signed-rank test for higher levels of contamination (see Hampel et al. 1986, Chapter 3).

20.5 Concluding Remarks

In this paper, we have studied a spatial sign test and a spatial signed-rank test for paired sample problems in infinite dimensional spaces. The tests are infinite dimensional extensions of the multivariate spatial sign and signed-rank tests considered earlier by Möttönen and Oja (1995), Möttönen et al. (1997), and Marden (1999) for finite dimensional data. We have shown that the asymptotic distributions of the proposed test statistics are Gaussian after appropriate centering and scaling. It is shown that both of the proposed tests are consistent for a class of alternatives that includes the standard location shift alternatives. It is observed that under suitable sequences of shrinking location shift alternatives, the spatial sign and signed-rank tests are asymptotically more powerful than some of the mean based paired sample tests for infinite dimensional data when the underlying distribution has heavy tails. Further, even for some infinite dimensional Gaussian distributions, the spatial sign and signed-rank tests considered in this paper are more powerful than some of the mean based tests for paired sample problems. The asymptotic results are corroborated by finite sample simulation results.

The proposed spatial sign and signed-rank test statistics can be computed very easily, and the associated tests can be implemented using their asymptotic Gaussian distributions. The covariance operators of the asymptotic Gaussian distributions can

be easily estimated as discussed in Sect. 20.2.1. For data in a Hilbert space, the implementations of the proposed tests become further simplified in view of the fact that in such a space, the squared norm of a Gaussian random element is distributed as a weighted sum of independent chi-square variables each with one degree of freedom.

The tests based on spatial signs and signed ranks studied in this paper do not require any moment assumption unlike some mean based tests for infinite dimensional data. These tests are also robust against contamination of the sample by outliers unlike the mean based tests considered in the paper. Between the two tests proposed in this paper, it is observed that the spatial sign test is better for some very heavy tailed distributions, while the spatial signed-rank test outperforms it in some cases when the distribution has lighter tails. Further, the spatial sign test is more robust than the spatial signed-rank test, when there is a large amount of contamination in the sample.

References

Araujo, A., Giné, E.: The Central Limit Theorem for Real and Banach Valued Random Variables. Wiley, New York/Chichester/Brisbane (1980)

Borovskikh, Y.V.: *UB*-statistics. Teor. Veroyatnost. i Primenen. **36**, 417–433 (1991)

Borwein, J.M., Vanderwerff, J.D.: Convex Functions: Constructions, Characterizations and Counterexamples. Cambridge University Press, Cambridge (2010)

Cardot, H., Cénac, P., Zitt, P.-A.: Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. Bernoulli **19**, 18–43 (2013)

Chakraborty, A., Chaudhuri, P.: A Wilcoxon-Mann-Whitney type test for infinite dimensional data. Biometrika **102**, 239–246 (2015)

Chakraborty, B., Chaudhuri, P., Oja, H.: Operating transformation retransformation on spatial median and angle test. Stat. Sin. **8**, 767–784 (1998)

Chaudhuri, P., Sengupta, D.: Sign tests in multidimension: inference based on the geometry of the data cloud. J. Am. Stat. Assoc. **88**, 1363–1370 (1993)

Cuevas, A., Febrero, M., Fraiman, R.: An anova test for functional data. Comput. Stat. Data Anal. **47**, 111–122 (2004)

Hallin, M., Paindaveine, D.: Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. Ann. Stat. **30**, 1103–1133 (2002)

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.: Robust Statistics: The Approach Based on Influence Functions. Wiley, New York (1986)

Hettmansperger, T.P., Nyblom, J., Oja, H.: Affine invariant multivariate one-sample sign tests. J. R. Stat. Soc. Ser. B **56**, 221–234 (1994)

Hettmansperger, T.P., Möttönen, J., Oja, H.: Affine-invariant multivariate one-sample signed-rank tests. J. Am. Stat. Assoc. **92**, 1591–1600 (1997)

Horváth, L., Kokoszka, P.: Inference for Functional Data with Applications. Springer, New York (2012)

Horváth, L., Kokoszka, P., Reeder, R.: Estimation of the mean of functional time series and a two-sample problem. J. R. Stat. Soc. Ser. B Stat. Methodol. **75**, 103–122 (2013)

Marden, J.I.: Multivariate rank tests. In: Ghosh, S. (ed.) Multivariate Analysis, Design of Experiments, and Survey Sampling, pp. 401–432. Dekker, New York (1999)

Möttönen, J., Oja, H.: Multivariate spatial sign and rank methods. J. Nonparametr. Stat. **5**, 201–213 (1995)

Möttönen, J., Oja, H., Tienari, J.: On the efficiency of multivariate spatial sign and rank tests. Ann. Stat. **25**, 542–552 (1997)

Oja, H.: Affine invariant multivariate sign and rank tests and corresponding estimates: a review. Scand. J. Stat. **26**, 319–343 (1999)

Oja, H.: Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks. Springer, New York (2010)

Peters, D., Randles, R.H.: A multivariate signed-rank test for the one-sample location problem. J. Am. Stat. Assoc. **85**, 552–557 (1990)

Puri, M.L., Sen, P.K.: Nonparametric Methods in Multivariate Analysis. Wiley, New York/London/Sydney (1971)

Randles, R.H.: A distribution-free multivariate sign test based on interdirections. J. Am. Stat. Assoc. **84**, 1045–1050 (1989)

Serfling, R.J.: Nonparametric multivariate descriptive measures based on spatial quantiles. J. Stat. Plann. Inference **123**, 259–278 (2004)

Vakhania, N.N., Tarieladze, V.I., Chobanyan, S.A.: Probability Distributions on Banach Spaces (Translated from the Russian and with a Preface by Wojbor A. Woyczynski). D. Reidel Publishing Co., Dordrecht (1987)

Yu, S., Tresp, V., Yu, K.: Robust multi-task learning with t-processes. In: Proceedings of the 24th International Conference on Machine Learning (Oregon, 2007), pp. 1103–1110. Omnipress, Corvallis, OR (2007)

Chapter 21
Semiparametric Analysis in Conditionally Independent Multivariate Mixture Models

Tracey W. Hammel, Thomas P. Hettmansperger, Denis H.Y. Leung, and Jing Qin

Abstract The conditional independence assumption is commonly used in multivariate mixture models in behavioral research. We propose an exponential tilt model to analyze data from a multivariate mixture distribution with conditionally independent components. In this model, the log ratio of the density functions of the components is modeled as a quadratic function in the observations. There are a number of advantages in this approach. First, except for the exponential tilt assumption, the marginal distributions of the observations can be completely arbitrary. Second, unlike some previous methods, which require the multivariate data to be discrete, modeling can be performed based on the original data.

Keywords Empirical likelihood • Exponential tilting • Repeated measures • Mixture distribution • Multivariate

21.1 Introduction

There are many applications where the interest is to classify n observations into m groups based on k measures on each observation. For example, Hettmansperger and Thomas (2000) and Cruz-Medina et al. (2004) described an experiment in developmental psychology where repeated measurements are made on children's responses to a cognitive task and the interest is to classify children into different groups based on the repeated measurements. The repeated measures data can

T.W. Hammel • T.P. Hettmansperger (✉)
Department of Statistics, Penn State University, University Park, PA, USA
e-mail: traceywrobel@gmail.com; tph@stat.psu.edu

D.H.Y. Leung
School of Economics, Singapore Management University, Singapore, Singapore
e-mail: denisleung@smu.edu.sg

J. Qin
Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA
e-mail: jingqin@niaid.nih.gov

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_21

be considered to come from a mixture of multivariate distributions, with the components corresponding to the response distributions in the different groups of observations, the number of components corresponding to the number of groups, and the mixing proportions corresponding to the proportions in the population that belong to the different groups. Two problems are of interest. First, to determine the number of groups. Second, to estimate the underlying component distributions and the mixing proportions.

Analysis of multivariate mixture distributions is a difficult problem (see, e.g., Titterington et al. 1985; Lindsay 1995; McLachlan and Peel 2000). Computation is commonly carried out using the EM algorithm (Dempster et al. 1977), which typically requires parametric distributional assumptions. However, a number of works (Thomas and Lohaus 1993; Hettmansperger and Thomas 2000; Hall and Zhou 2003; Cruz-Medina et al. 2004; Leung and Qin 2006; Chang and Walther 2007; Benaglia et al. 2009) showed that a semiparametric or nonparametric approach might be a flexible and robust alternative to a parametric approach.

In the situation described in the first paragraph, each child who participated in the study was given a total of six tasks, each randomly selected from a large pool of similar tasks. As a result, it is unlikely for a child to predict the next task and hence the responses to different tasks can be considered independent of each other. This observation led us to make the assumption of conditional independence, which means that conditional on component membership, the multivariate component distribution is the product of its marginals; see also Sect. 21.7. Under the conditional independence assumption, the m component mixture has probability density function (PDF) or probability mass function (PMF)

$$h(x_1, \ldots, x_k) = \sum_{l=1}^{m} \lambda_l \prod_{j=1}^{k} f_{lj}(x_j), \tag{21.1}$$

where λ_l is the mixing proportion for the lth component and f_{lj} is the PDF (or PMF) for the lth component of the jth repeated measure. Later, we impose further structural assumptions. Unlike previous works (Hettmansperger and Thomas 2000; Cruz-Medina et al. 2004; Leung and Qin 2006; Chang and Walther 2007), (21.1) does not require identical marginal distributions. Conditional independence of multivariate data can also be seen as a special case of the popular random effects model with clustered data (Liu and Pierce 1994; Qu and Hadgu 1998).

A histogram of the data in the study (Cruz-Medina et al. 2004, Fig. 1) shows that the data distribution is unremarkable; there is no immediate resemblance to any well-known distribution. This observation motivates a semiparametric approach to analyzing the data. We assume the component densities are related by an exponential tilt (density-ratio) model (Anderson 1979). For a two-component mixture with PDFs f and g, our exponential tilt model assumes f and g are related by $\log\left(g(x)/f(x)\right) = \alpha + \beta x + \gamma x^2$. As a parallel to the Cox proportional hazards model and the Lehmann alternative model, the exponential tilt model is very versatile, due to its natural connection to the logistic model. Kay and Little (1986) discuss various versions

of the exponential tilt model for some common distributions. Because the normal PDF has a quadratic exponent, any two normal PDFs satisfy the condition for the exponential tilt model described above. In many situations where common parametric distributions do not fit the observed data well, the exponential tilt model still can provide excellent fits (Qin and Zhang 1997; Nagelkerke et al. 2001; Zhang 2001; Qin et al. 2002; White and Thompson 2003). Efron and Tibshirani (1996) argue that the exponential tilt is a favorable compromise between parametric and nonparametric density estimation.

The rest of this paper is organized as follows. Details of the method are described in Sect. 21.2. The exponential tilt model is formulated using an empirical likelihood (Owen 1988). Under mild conditions, the model is uniquely identifiable up to label switching, which is important for estimating the underlying mixture structure. In Sect. 21.3, we present an EM algorithm. Estimation of features of the component distributions is discussed in Sect. 21.4. In Sect. 21.5, we evaluate the method using simulations. In Sect. 21.6, we propose a model selection criterion based on the BIC (Bayesian Information Criterion; Schwarz 1978) to estimate the number of components in the mixture. In Sect. 21.7, the method is applied to the data of Cruz-Medina et al. (2004). Section 21.8 concludes with a discussion of possible future work.

21.2 Exponential Tilt Model

We consider n multivariate vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from an m component, k dimensional multivariate mixture distribution, where $\mathbf{X}_i^\top = (x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$.

Let $(x_1, \ldots, x_k)$ be a generic observation, then its joint PDF can be written as

$$h(x_1, \ldots, x_k) = \lambda_1 \prod_{j=1}^{k} f_j(x_j) + \sum_{l=2}^{m} \lambda_l \prod_{j=1}^{k} g_{lj}(x_j), \tag{21.2}$$

where f_j and g_{lj} represent univariate PDFs, λ_1 is the mixing proportion of component one (the baseline distribution), $0 < \lambda_l < 1$ is the mixing proportion of component l and $\sum_{l=1}^{m} \lambda_l = 1$. Let H, F_j, and G_{lj} denote the CDFs corresponding to h, f_j, and g_{lj}, respectively.

Let f_j and g_{lj} be related by a quadratic exponential tilt model

$$\log\left(g_{lj}(x_j)/f_j(x_j)\right) = \alpha_{lj} + \beta_{lj} x_j + \gamma_{lj} x_j^2, \tag{21.3}$$

where α_{lj}, β_{lj}, and γ_{lj} are unknown parameters. The PDF (21.2) can be re-written as

$$h(x_1, \ldots, x_k) = \left[\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left\{\sum_{j=1}^{k} \alpha_{lj} + \beta_{lj} x_j + \gamma_{lj} x_j^2\right\}\right] \prod_{j=1}^{k} f_j(x_j). \tag{21.4}$$

Theorem 8 of Allman et al. (2009) states that a mixture of the form (1) is uniquely identifiable up to label switching provided that $k \geq 3$ and, for each $j = 1, \ldots, k$, the m distributions are linearly independent. This result makes sense since linear independence precludes expressing any one of the coordinate distributions as a linear combination of the other $m-1$ distributions. Since, in our case, $\sum_{l=1}^{m} \lambda_l = 1$ and $0 < \lambda_l < 1$, and for each $j = 1, \ldots, k$

$$\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left\{\alpha_{lj} + \beta_{lj} x_j + \gamma_{lj} x_j^2\right\} \neq 0 \quad \text{for} \quad -\infty < x_j < \infty,$$

identifiability follows for model (21.4). For earlier results on identifiability in nonparametric mixtures, see Hall and Zhou (2003), Hall et al. (2005), and Elmore et al. (2005).

Let $\boldsymbol{\theta}_{lj}^{\top} = (\alpha_{lj}, \beta_{lj}, \gamma_{lj})$, $\tilde{\mathbf{x}}_{ij}^{\top} = (1, x_{ij}, x_{ij}^2)$, $\tilde{\mathbf{x}}_j$ the counterpart of $\tilde{\mathbf{x}}_{ij}$ for a generic observation, $\boldsymbol{\lambda}^{\top} = (\lambda_1, \ldots, \lambda_m)$, $\boldsymbol{\theta}^{\top} = (\boldsymbol{\theta}_{21}, \ldots, \boldsymbol{\theta}_{mk})$ and $\boldsymbol{\delta}^{\top} = (\boldsymbol{\lambda}^{\top}, \boldsymbol{\theta}^{\top})$, then the likelihood based on the observed data is

$$L(\boldsymbol{\delta}, F_1, \ldots, F_k) = \prod_{i=1}^{n} \left[\left\{ \lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^{\top} \boldsymbol{\theta}_{lj} \right) \right\} \prod_{j=1}^{k} dF_j(x_{ij}) \right].$$

The maximizing F_j only jumps at each observed x_{ij} (Owen 1988). Let the jump sizes be p_{ij}, then the log-likelihood is

$$\ell(\boldsymbol{\delta}, p_{11}, \ldots, p_{nk}) = \sum_{i=1}^{n} \left[\log \left\{ \lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^{\top} \boldsymbol{\theta}_{lj} \right) \right\} + \sum_{j=1}^{k} \log p_{ij} \right]. \tag{21.5}$$

For fixed $\boldsymbol{\delta}$, ℓ can be maximized with respect to the p_{ij}s subject to the constraints

$$\sum_{i=1}^{n} p_{ij} = 1, \;\; p_{ij} \geq 0, \;\; \sum_{i=1}^{n} p_{ij} \exp\left(\tilde{\mathbf{x}}_{ij}^{\top} \boldsymbol{\theta}_{lj}\right) = 1, \;\; j = 1, \ldots, k, l = 2, \ldots, m. \tag{21.6}$$

The last k constraints in (21.6) come from model (21.3) and are responsible for ensuring that the resulting g_{lj} are proper PDFs. The constrained maximization can be accomplished using a Lagrange multiplier argument, which leads to

$$p_{ij} = \frac{1}{n} \left[\frac{1}{1 + \frac{1}{n} \sum_{l=2}^{m} \eta_{lj} \{\exp(\tilde{\mathbf{x}}_{ij}^{\top} \boldsymbol{\theta}_{lj}) - 1\}} \right], \;\; i = 1, \ldots, n, j = 1, \ldots, k, \tag{21.7}$$

where $\boldsymbol{\eta}^\top \equiv (\eta_{21}, \ldots, \eta_{mk})$ are Lagrange multipliers determined by

$$\sum_{i=1}^{n} \frac{\exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1}{1 + \frac{1}{n}\sum_{l=2}^{m} \eta_{lj}\{\exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1\}} = 0, \quad j = 1, \ldots, k, l = 2, \ldots, m. \tag{21.8}$$

Note that if the exponential tilt parameters $\boldsymbol{\theta}_{lj}^\top = \mathbf{0}$, then (21.7) would simply be the weights found for the empirical distribution, namely $1/n$. Substituting the p_{ij}s back into (21.5) gives a log-profile likelihood

$$\ell_p(\boldsymbol{\delta}) = \sum_{i=1}^{n} \log \left\{ \lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj} \right) \right\} - \sum_{i=1}^{n} \sum_{j=1}^{k} \log \left[n + \sum_{l=2}^{m} \eta_{lj}\{\exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1\} \right]. \tag{21.9}$$

Denote the maximum semiparametric likelihood estimate obtained from maximizing $\ell_p(\boldsymbol{\delta})$ by $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\eta}}$ the corresponding value of the Lagrange multipliers at the maximum likelihood. The following theorem describes the large sample behavior of the maximum semiparametric likelihood estimate.

Theorem 21.1 *Let* $\mathbf{U}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\eta}) = (u_1, u_2, u_3)$, *where* $u_1(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\eta}) = \partial \ell_p / \partial \boldsymbol{\theta}_{lj}$, $u_2(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\eta}) = \partial \ell_p / \partial \eta_{lj}$, $u_3(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\eta}) = \partial \ell_p / \partial \lambda_l$. *Let* $\boldsymbol{\delta}^0 \equiv (\boldsymbol{\lambda}^0, \boldsymbol{\theta}^0, \boldsymbol{\eta}^0)$ *be the true values of* $\boldsymbol{\delta} \equiv (\boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{\eta})$ *and let the superscript "0" denote a quantity evaluated at* $\boldsymbol{\delta}^0$. *Assume the conditions hold:*

[C1] $E\{\mathbf{U}^0(\mathbf{U}^0)^\top\}$ *is positive definite; and the rank of* $E(\partial \mathbf{U}^0 / \partial \boldsymbol{\delta})$ *is* $2(m-1)k + (m-1)$, *which is also the dimension of* $\boldsymbol{\delta}$.

[C2] $\partial^2 \mathbf{U}(\boldsymbol{\delta}) / (\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^\top)$ *is continuous in a neighborhood of* $\boldsymbol{\delta}^0$ *where* $\|\partial \mathbf{U}(\boldsymbol{\delta}) / \partial \boldsymbol{\delta}\|$ *is bounded,* $E(||\mathbf{U}(\boldsymbol{\delta})||)^2 < \infty$.

[C3] Functions are sufficiently smooth to allow differentiation under the integral. and $0 < \lambda_1, \ldots, \lambda_m < 1$, *then for any sufficiently smooth function g,*

$$\sqrt{n} g(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0, \hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0, \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_g).$$

Furthermore, asymptotically, the estimates achieve semiparametric efficiency.

Proof *For a matrix* $\mathbf{a}$, *denote its i, jth element by* $a_{i,j}$ *and let* $\mathbf{A} = E(\mathbf{a})$ *where the expectation is taken under* $\boldsymbol{\delta}^0$. *Write* $w_{il}^- = \lambda_l / \{\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj})\}$, $v_{ilj}^- = \eta_{lj} / [1 + 1/n \sum_{l=2}^{m} \eta_{lj}\{\exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1\}]$. *Let* $\dot{\mathbf{w}}_{il,\boldsymbol{\theta}_{l'j'}}^0 = w_{il}^{-0} \partial / \partial \boldsymbol{\theta}_{l'j'} \{\lambda_l^0$

$\exp(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^{\top} \boldsymbol{\theta}_{lj}^{0})\}$, $\dot{\mathbf{v}}_{ilj,\boldsymbol{\theta}_{l'j'}}^{0} = v_{lj}^{-0} \partial / \partial \boldsymbol{\theta}_{l'j'} [\boldsymbol{\eta}_{lj}^{0} \{\exp(\tilde{\mathbf{x}}_{ij}^{\top} \boldsymbol{\theta}_{lj}^{0}) - 1\}]$ *and similarly for* $\dot{\mathbf{w}}_{il,\eta_{l'j'}}^{0}$, $\dot{\mathbf{w}}_{il,\lambda_{l'}}^{0}$, $\dot{\mathbf{v}}_{ilj,\eta_{l'j'}}^{0}$, $\dot{\mathbf{v}}_{ilj,\lambda_{l'}}^{0}$, $\dot{\mathbf{v}}_{l,\boldsymbol{\theta}_{l'j'}}^{-0}$, *and* $\dot{\mathbf{v}}_{l,\eta_{l'j'}}^{-0}$.

$$\frac{\partial \mathbf{U}(\boldsymbol{\theta}^{0}, \boldsymbol{\lambda}^{0}, \boldsymbol{\eta}^{0})}{\partial (\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\eta})^{\top}} = \mathbf{a} = \begin{pmatrix} \mathbf{a}_{11} & \mathbf{a}_{12} & \mathbf{a}_{13} \\ \mathbf{a}_{21} & \mathbf{a}_{22} & \mathbf{0} \\ \mathbf{a}_{31} & \mathbf{0} & \mathbf{a}_{33} \end{pmatrix} \to \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{0} \\ \mathbf{A}_{31} & \mathbf{0} & \mathbf{A}_{33} \end{pmatrix} = \mathbf{A}, \tag{21.10}$$

where

$$\begin{aligned}
\mathbf{a}_{11}(lj, l'j') &= \frac{\partial u_1^0}{\partial \boldsymbol{\theta}_{l'j'}} = \sum_{i=1}^{n} \Big[\dot{\mathbf{w}}_{il,\boldsymbol{\theta}_{l'j'}}^{0} - w_{il}^{0} \dot{\mathbf{w}}_{il',\boldsymbol{\theta}_{l'j'}}^{0} \\
&\qquad - \Big\{ \dot{\mathbf{v}}_{ilj,\boldsymbol{\theta}_{l'j'}}^{0} - v_{ilj}^{0} (\dot{\mathbf{v}}_{ilj,\boldsymbol{\theta}_{l'j'}}^{0} - \dot{\mathbf{v}}_{ilj,\boldsymbol{\theta}_{l'j'}}^{-0}) \Big\} \Big] \tilde{\mathbf{x}}_{ij}^{\top}, \\
\mathbf{a}_{12}(lj, l'j') &= \frac{\partial u_1^0}{\partial \eta_{l'j'}} = \sum_{i=1}^{n} \dot{\mathbf{v}}_{ilj,\eta_{l'j'}}^{0} - v_{ilj}^{0} (\dot{\mathbf{v}}_{ilj,\eta_{l'j'}}^{0} - \dot{\mathbf{v}}_{ilj,\eta_{l'j'}}^{-0}), \\
\mathbf{a}_{13}(l, l') &= \frac{\partial u_1^0}{\partial \lambda_{l'}} = \sum_{i=1}^{n} \dot{\mathbf{w}}_{il,\lambda_{l'}}^{0} - w_{il}^{0} \dot{\mathbf{w}}_{il',\lambda_{l'}}^{0}, \\
\mathbf{a}_{21}(lj, l'j') &= \frac{\partial u_2^0}{\partial \boldsymbol{\theta}_{l'j'}} = \sum_{i=1}^{n} \frac{1}{\eta_{lj}} \Big\{ \dot{\mathbf{v}}_{ilj,\boldsymbol{\theta}_{l'j'}}^{0} - (v_{ilj}^{0} - v_{ilj}^{-0}) (\dot{\mathbf{v}}_{ilj,\boldsymbol{\theta}_{l'j'}}^{0} - \dot{\mathbf{v}}_{ilj,\boldsymbol{\theta}_{l'j'}}^{-0}) \Big\}, \\
\mathbf{a}_{22}(lj, l'j') &= \frac{\partial u_2^0}{\partial \eta_{l'j'}} = \sum_{i=1}^{n} \frac{1}{\eta_{lj}} \Big\{ \dot{\mathbf{v}}_{ilj,\eta_{l'j'}}^{0} - (v_{ilj}^{0} - v_{ilj}^{-0}) (\dot{\mathbf{v}}_{ilj,\eta_{l'j'}}^{0} - \dot{\mathbf{v}}_{ilj,\eta_{l'j'}}^{-0}) \Big\}, \\
\mathbf{a}_{31}(lj, l'j') &= \frac{\partial u_3^0}{\partial \boldsymbol{\theta}_{l'j'}} = \sum_{i=1}^{n} \frac{1}{\lambda_{l}} \Big\{ \dot{\mathbf{w}}_{il,\boldsymbol{\theta}_{l'j'}}^{0} - (w_{il}^{0} - w_{il}^{-0}) \dot{\mathbf{w}}_{il',\boldsymbol{\theta}_{l'j'}}^{0} \Big\}, \\
\mathbf{a}_{33}(l, l') &= \frac{\partial u_3^0}{\partial \lambda_{l'}} = \sum_{i=1}^{n} \frac{1}{\lambda_{l}} \Big\{ \dot{\mathbf{w}}_{il,\boldsymbol{\theta}_{l'j'}}^{0} - (w_{il}^{0} - w_{il}^{-0}) \dot{\mathbf{w}}_{il',\boldsymbol{\theta}_{l'j'}}^{0} \Big\}.
\end{aligned}$$

Define row vectors $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ *such that* $\mathbf{b}_1(lj) = \sum_{i=1}^{n} (w_{il}^{0} - v_{ilj}^{0}) \tilde{\mathbf{x}}_{ij}^{\top}$, $\mathbf{b}_2(lj) = \sum_{i=1}^{n} \frac{1}{\eta_{lj}} (v_{ilj}^{0} - v_{ilj}^{-0})$, $\mathbf{b}_3(l) = \sum_{i=1}^{n} \frac{1}{\lambda_{l}} (w_{il}^{0} - w_{il}^{-0})$. *Then* $\sqrt{n} \mathbf{U}(\boldsymbol{\theta}^{0}, \boldsymbol{\lambda}^{0}, \boldsymbol{\eta}^{0}) = n^{-1} (\mathbf{b}_1^{\top}, \mathbf{b}_2^{\top}, \mathbf{b}_3^{\top})^{\top} \xrightarrow{d} N(\mathbf{0}, \mathbf{W})$, *where*

$$\mathbf{W} = E \begin{pmatrix} \mathbf{b}_1^{\top} \mathbf{b}_1 & \mathbf{b}_1^{\top} \mathbf{b}_2 & \mathbf{b}_1^{\top} \mathbf{b}_3 \\ \mathbf{b}_2^{\top} \mathbf{b}_1 & \mathbf{b}_2^{\top} \mathbf{b}_2 & \mathbf{b}_2^{\top} \mathbf{b}_3 \\ \mathbf{b}_3^{\top} \mathbf{b}_1 & \mathbf{b}_3^{\top} \mathbf{b}_2 & \mathbf{b}_3^{\top} \mathbf{b}_3 \end{pmatrix} = \begin{pmatrix} \mathbf{W}_1 & \mathbf{W}_2 \\ \mathbf{W}_2^{\top} & \mathbf{W}_3 \end{pmatrix}.$$

It then follows that

$$\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0\\ \hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}^0\\ \hat{\boldsymbol{\eta}}-\boldsymbol{\eta}^0\end{pmatrix}=n^{-1}(\mathbf{a}^\top)^{-1}(\mathbf{b}_1^\top,\mathbf{b}_2^\top,\mathbf{b}_3^\top)^\top+o_p(\sqrt{n})\xrightarrow{d}N\left\{\mathbf{0},(\mathbf{A}^T)^{-1}\mathbf{W}\mathbf{A}^{-1}\right\}.$$

We can also study the behavior of particular parameters of interest. In particular,

$$\sqrt{n}(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}^0)\xrightarrow{d}N(\mathbf{0},\boldsymbol{\Sigma}\equiv\mathbf{C}^\top\mathbf{W}_{11}\mathbf{C}+\mathbf{D}\mathbf{W}_{12}^\top\mathbf{C}+\mathbf{C}^\top\mathbf{W}_{12}\mathbf{C}+\mathbf{D}\mathbf{W}_{22}\mathbf{D}),$$

where

$$\begin{pmatrix}\mathbf{V}_{11}&\mathbf{V}_{12}&\mathbf{V}_{13}\\ \mathbf{V}_{21}&\mathbf{V}_{22}&\mathbf{V}_{23}\\ \mathbf{V}_{31}&\mathbf{V}_{23}&\mathbf{V}_{33}\end{pmatrix}=\mathbf{A}^{-1},\ \ \mathbf{D}=\left(\mathbf{A}_{33}-\mathbf{A}_{13}^\top\mathbf{V}_{11}\mathbf{A}_{13}\right),\ \ \mathbf{C}=\begin{pmatrix}\mathbf{V}_{11}\mathbf{A}_{13}\\ \mathbf{V}_{21}\mathbf{A}_{13}\end{pmatrix}\mathbf{D}.$$

We now prove the semiparametric efficiency of the proposed method. Write $\psi(x_1,\ldots,x_k,\boldsymbol{\delta})=\lambda_1+\sum_{l=2}^m\lambda_l\exp(\sum_{j=1}^k\tilde{\mathbf{x}}_j^\top\boldsymbol{\theta}_{lj})$. *For a finite dimensional parameter* $\boldsymbol{\phi}$, *consider a parametric submodel of* $h(x_1,\ldots,x_k)$

$$h(x_1,\ldots,x_k,\boldsymbol{\delta},\boldsymbol{\phi})=\psi(x_1,\ldots,x_k,\boldsymbol{\delta})\prod_{j=1}^k f_j(x_j,\boldsymbol{\phi}).\tag{21.11}$$

The profile likelihood $L_p(\boldsymbol{\delta})$ *is proportional to*

$$\frac{h(x_1,\ldots,x_k,\boldsymbol{\delta},\boldsymbol{\phi})}{h_1(x_1,\boldsymbol{\delta},\boldsymbol{\phi})\cdots h_k(x_k,\boldsymbol{\delta},\boldsymbol{\phi})},\tag{21.12}$$

where $h_j(x_j,\boldsymbol{\delta},\boldsymbol{\phi})=\{\lambda_1+\sum_{l=2}^m\lambda_l\exp(\tilde{\mathbf{x}}_j^\top\boldsymbol{\theta}_{lj})\}f_j(x_j,\boldsymbol{\phi})$. *Let* $\mathbf{S}_{\boldsymbol{\delta}},\mathbf{S}_{\boldsymbol{\phi}}$ *be the score functions of* $\boldsymbol{\delta}$ *and* $\boldsymbol{\phi}$ *based on (21.11) and (21.12). Write* $\psi(\boldsymbol{\delta})=\psi(x_1,\ldots,x_k,\boldsymbol{\delta})$, $h(\boldsymbol{\delta},\boldsymbol{\phi})=h(x_1,\ldots,x_k,\boldsymbol{\delta},\boldsymbol{\phi})$ $h_j(\boldsymbol{\delta},\boldsymbol{\phi})=h_j(x_j,\boldsymbol{\delta},\boldsymbol{\phi})$, $f_j(\boldsymbol{\phi})=f_j(x_j,\boldsymbol{\phi})$, $\dot{\mathbf{h}}_{\boldsymbol{\delta}}(\boldsymbol{\delta},\boldsymbol{\phi})=\partial h/\partial\boldsymbol{\delta}$, $\dot{\_}_{\boldsymbol{\delta}}(\boldsymbol{\delta})=\partial\psi/\partial\boldsymbol{\delta}$ *and* $\dot{\mathbf{f}}_{\boldsymbol{\phi}}(\boldsymbol{\phi})=\partial f/\partial\boldsymbol{\phi}$. *Then*

$$\mathbf{S}_{\boldsymbol{\delta}}=\frac{\dot{\mathbf{h}}_{\boldsymbol{\delta}}(\boldsymbol{\delta},\boldsymbol{\phi})}{h(\boldsymbol{\delta},\boldsymbol{\phi})}-\sum_{j=1}^k\frac{\dot{\mathbf{h}}_{j,\boldsymbol{\delta}}(\boldsymbol{\delta},\boldsymbol{\phi})}{h_j(\boldsymbol{\delta},\boldsymbol{\phi})}=\mathbf{S}_{\boldsymbol{\delta}}^A+\sum_{j=1}^k\mathbf{S}_{j,\boldsymbol{\delta}}^B,$$

$$\mathbf{S}_{\boldsymbol{\phi}}=\frac{\dot{\mathbf{h}}_{\boldsymbol{\delta}}(\boldsymbol{\delta},\boldsymbol{\phi})}{h(\boldsymbol{\delta},\boldsymbol{\phi})}=-\sum_{j=1}^k\frac{\dot{\mathbf{f}}_{j,\boldsymbol{\phi}}(\boldsymbol{\phi})}{f_j(\boldsymbol{\phi})}=\sum_{j=1}^k\mathbf{S}_{j,\boldsymbol{\phi}}.$$

We will show that $\mathbf{S}_{\boldsymbol{\delta}}$ *and* $\mathbf{S}_{\boldsymbol{\phi}}$ *are orthogonal by showing* $E(\mathbf{S}^A_{\boldsymbol{\delta}}\mathbf{S}_{j,\boldsymbol{\phi}}) = \mathbf{0}$ *and* $E(\mathbf{S}^B_{j,\boldsymbol{\delta}}\mathbf{S}_{j',\boldsymbol{\phi}}) = \mathbf{0}, j, j' = 1, \ldots, k$ *and hence, the estimator is efficient. Denote* $\int \cdot d\mathbf{x} \equiv \int \cdots \int \cdot dx_1 \cdots dx_k$ *and* $\int \cdot d\mathbf{x}_{-1} \equiv \int \cdots \int \cdot dx_2 \cdots dx_k$.

$$
\begin{aligned}
E(\mathbf{S}^A_{\boldsymbol{\delta}}\mathbf{S}_{j,\boldsymbol{\phi}}) &= \int \frac{\dot{\mathbf{h}}_{\boldsymbol{\delta}}(\boldsymbol{\delta},\boldsymbol{\phi})}{h(\boldsymbol{\delta},\boldsymbol{\phi})} \frac{\dot{\mathbf{f}}_{j,\boldsymbol{\phi}}(\boldsymbol{\phi})}{f_j(\boldsymbol{\phi})} h(\boldsymbol{\delta},\boldsymbol{\phi}) d\mathbf{x} \\
&= \int \dot{\llcorner}_{\boldsymbol{\delta}}(\boldsymbol{\delta}) \prod_{j=1}^{k} f_j(\boldsymbol{\phi}) \frac{\dot{\mathbf{f}}_{j,\boldsymbol{\phi}}(\boldsymbol{\phi})}{f_j(\boldsymbol{\phi})} d\mathbf{x} \\
&= \frac{\partial}{\partial \boldsymbol{\delta}} \int \frac{\partial}{\partial \boldsymbol{\phi}} \{\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp(\sum_{j=2}^{k} \tilde{\mathbf{x}}_j^\top \boldsymbol{\theta}_{lj})\} \prod_{j=2}^{k} f_j(\boldsymbol{\phi}) d\mathbf{x}_{-1} = \mathbf{0}. \\
E(\mathbf{S}^B_{j,\boldsymbol{\delta}}\mathbf{S}_{j',\boldsymbol{\phi}}) &= \int \frac{\dot{\mathbf{h}}_{j,\boldsymbol{\delta}}(\boldsymbol{\delta},\boldsymbol{\phi})}{h_j(\boldsymbol{\delta},\boldsymbol{\phi})} \frac{\dot{\mathbf{f}}_{j',\boldsymbol{\phi}}(\boldsymbol{\phi})}{f_{j'}(\boldsymbol{\phi})} \psi(\boldsymbol{\delta}) \prod_{j=1}^{k} f_j(\boldsymbol{\phi}) d\mathbf{x} \\
&= \int \frac{\dot{\mathbf{h}}_{j,\boldsymbol{\delta}}(\boldsymbol{\delta},\boldsymbol{\phi})}{h_j(\boldsymbol{\delta},\boldsymbol{\phi})} \frac{\partial}{\partial \boldsymbol{\phi}} \{\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp(\sum_{j \neq j'} \tilde{\mathbf{x}}_j^\top \boldsymbol{\theta}_{lj})\} \prod_{j \neq j'}^{k} f_j(\boldsymbol{\phi}) d\mathbf{x}_{-1} = \mathbf{0}.
\end{aligned}
$$

□

The theorem allows us to draw inference about the mixture parameter $\boldsymbol{\lambda}$, as well as other quantities, such as component moments, that are smooth functions of the distribution parameters.

21.3 Estimation

Estimates of the parameters can be found by differentiating (21.9) with respect to each of the parameters and solving the simultaneous equations:

$$
\frac{\partial \ell_p}{\partial \boldsymbol{\theta}_{lj}} \Rightarrow \sum_{i=1}^{n} w_{il} \tilde{\mathbf{x}}_{ij}^\top - \sum_{i=1}^{n} v_{ilj} \tilde{\mathbf{x}}_{ij}^\top = 0, \tag{21.13}
$$

$$
\frac{\partial \ell_p}{\partial \eta_{lj}} \Rightarrow \sum_{i=1}^{n} \frac{\frac{1}{n}\{\exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1\}}{1 + \frac{1}{n}\sum_{l=2}^{m} \eta_{lj}\{\exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1\}} = 0, \tag{21.14}
$$

$$
\frac{\partial \ell_p}{\partial \lambda_l} \Rightarrow \sum_{i=1}^{n} \frac{\exp(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1}{\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj})} = 0, \tag{21.15}
$$

for $l = 2, \ldots, m$ and $j = 1, \ldots, k$ and w_{il} and v_{ilj} are defined by

$$w_{il} = \frac{\lambda_l \exp(\sum_{j=1}^k \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj})}{\lambda_1 + \sum_{l=2}^m \lambda_l \exp(\sum_{j=1}^k \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj})},$$
$$v_{ilj} = \frac{\frac{1}{n}\eta_{lj} \exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj})}{1 + \frac{1}{n}\sum_{l=2}^m \eta_{lj}\{\exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1\}}. \tag{21.16}$$

Notice that in (21.14), we have used the fact that $\lambda_1 = 1 - \sum_{l=2}^m \lambda_l$. There is no explicit solution for λ_l, $l = 1, \ldots, k$. Therefore, we develop an EM type algorithm (Dempster et al. 1977). Define the latent variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ where $\mathbf{Z}_i^\top = (z_{i1}, \ldots, z_{im})$ for the component membership for the ith observation in the data set. If the ith observation belonged to the lth component, then $\mathbf{Z}_i^\top$ is a vector of $m - 1$ 0s and a single 1 in the lth position. Furthermore, $\sum_{l=1}^m z_{il} = 1$. Of course, $\mathbf{Z}_i, i = 1, \ldots, n$ are not observed. We define the "complete data" as $\{(\mathbf{X}_1, \mathbf{Z}_1), \ldots, (\mathbf{X}_n, \mathbf{Z}_n)\}$; then, a complete data semiparametric likelihood is

$$\begin{aligned} L_c(\boldsymbol{\delta}, F_1, \ldots, F_k) = \prod_{i=1}^n \Bigg[& \left\{ \lambda_1 \prod_{j=1}^k F_j(x_{ij}) \right\}^{z_{i1}} \\ & \prod_{l=2}^m \left\{ \lambda_l \exp\left(\sum_{j=1}^k \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj} \right) \prod_{j=1}^k dF_j(x_{ij}) \right\}^{z_{il}} \Bigg] \\ = \prod_{i=1}^n \Bigg[& \lambda_1^{z_{i1}} \prod_{l=2}^m \lambda_l^{z_{il}} \exp\left(z_{il} \sum_{j=1}^k \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj} \right) \prod_{j=1}^k dF_j(x_{ij})^{z_{il}} \Bigg]. \end{aligned}$$

Using p_{ij} as the jump size of F_j at x_{ij}, the complete data log-likelihood is

$$\begin{aligned} & \ell_c(\boldsymbol{\delta}, p_{ij}, i = 1, \ldots, n, j = 1, \ldots, k) \\ & = \sum_{i=1}^n \sum_{1=1}^m z_{il} \log \lambda_l + \sum_{i=1}^n \sum_{1=2}^m z_{il} \sum_{j=1}^k \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj} + \sum_{i=1}^n \sum_{j=1}^k \log p_{ij}, \end{aligned} \tag{21.17}$$

where (21.6) still hold and p_{ij}s can be profiled out using (21.7) and (21.8).

Let the parameter estimates at iteration t of the EM algorithm be $[\boldsymbol{\delta}^{(t)}]^\top = (\boldsymbol{\theta}_{21}^{(t)}, \ldots, \boldsymbol{\theta}_{mk}^{(t)}, \lambda_1^{(t)}, \ldots, \lambda_m^{(t)})$ and write

$$w_{il}^{(t)} = \mathrm{E}(z_{il} | \boldsymbol{\delta}^{(t)}, x_1, \ldots, x_n) = \frac{\lambda_l^{(t)} \exp(\sum_{j=1}^k \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}^{(t)})}{\lambda_1^{(t)} + \sum_{l=2}^m \lambda_l^{(t)} \exp(\sum_{j=1}^k \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}^{(t)})}. \tag{21.18}$$

Since (21.17) is linear in z_{il}s, substituting (21.18) for z_{il}s and (21.7) for p_{ij}s in (21.17) gives the expected complete data profile log-likelihood (E-step) at iteration $t+1$,

$$Q(\boldsymbol{\delta}, \boldsymbol{\delta}^{(t)}) = \mathrm{E}(\ell_c|\boldsymbol{\delta}^{(t)}, x_1, \ldots, x_n) = \sum_{i=1}^{n}\sum_{1=1}^{m} w_{il}^{(t)} \log \lambda_l + \sum_{i=1}^{n}\sum_{1=2}^{m} w_{il}^{(t)} \sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^{\top} \boldsymbol{\theta}_{lj}$$
$$- \sum_{i=1}^{n}\sum_{j=1}^{k} \log[n + \sum_{l=2}^{m} \eta_{lj}\{\exp(\tilde{\mathbf{x}}_{ij}^{\top}\boldsymbol{\theta}_{lj}) - 1\}].$$

The M-step maximizes $Q(\boldsymbol{\delta}, \boldsymbol{\delta}^{(t)})$ with respect to $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$. Since $\lambda_l, l = 1, \ldots, m$, satisfy the constraint $\sum_{l=1}^{m} \lambda_l = 1$, we immediately obtain

$$\hat{\lambda}_l^{(t+1)} = \frac{\sum_{i=1}^{n} w_{il}^{(t)}}{n}. \tag{21.19}$$

Differentiating $Q(\boldsymbol{\delta}, \boldsymbol{\delta}^{(t)})$ with respect to the other parameters gives exactly the same equations as (21.13) to (21.14), but with $w_{il}^{(t)}$s replacing w_{il}s. Using (21.19) and replacing w_{il}s by $w_{il}^{(t)}$s in (21.13) and (21.14) now gives

$$n\left(\frac{\eta_{lj}^{(t+1)}}{n}\right) - \sum_{i=1}^{n} w_{il}^{(t)} = 0 \Rightarrow \frac{\eta_{lj}^{(t+1)}}{n} = \frac{\sum_{i=1}^{n} w_{il}^{(t)}}{n} = \lambda_l^{(t+1)}. \tag{21.20}$$

Using (21.20) in (21.13) now gives

$$\sum_{i=1}^{n} \frac{\lambda_l^{(t)} \exp(\tilde{\mathbf{x}}_{ij}^{\top}\boldsymbol{\theta}_{lj}^{(t)})\tilde{\mathbf{x}}_{ij}^{\top}}{\lambda_1^{(t)} + \sum_{l=2}^{m} \lambda_l^{(t)} \exp(\tilde{\mathbf{x}}_{ij}^{\top}\boldsymbol{\theta}_{lj}^{(t)})} - \frac{\lambda_l^{(t+1)} \exp(\tilde{\mathbf{x}}_{ij}^{\top}\boldsymbol{\theta}_{lj})\tilde{\mathbf{x}}_{ij}^{\top}}{\lambda_1^{(t+1)} + \sum_{l=2}^{m} \lambda_l^{(t+1)} \exp(\tilde{\mathbf{x}}_{ij}^{\top}\boldsymbol{\theta}_{lj})} = 0, \tag{21.21}$$

which can be used to easily solve for $\boldsymbol{\theta}_{lj}^{(t+1)}$ by a Newton-Raphson procedure.

To show that our EM algorithm increases $\ell_p(\boldsymbol{\delta})$ at every step, we note that

$$Q(\boldsymbol{\delta}, \boldsymbol{\delta}^{(t)}) \leq \sum_{i=1}^{n} \log\left\{w_{i1}^{(t)}\lambda_1 + \sum_{1=2}^{m} w_{il}^{(t)}\lambda_l \exp\left(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^{\top}\boldsymbol{\theta}_{lj}\right)\right\} + \sum_{i=1}^{n}\sum_{j=1}^{k} \log p_{ij}$$
$$< \sum_{i=1}^{n} \log\left\{\lambda_1 + \sum_{1=2}^{m} \lambda_l \exp\left(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^{\top}\boldsymbol{\theta}_{lj}\right)\right\} + \sum_{i=1}^{n}\sum_{j=1}^{k} \log p_{ij} = \ell_p(\boldsymbol{\delta}),$$

Since $\boldsymbol{\delta}^{(t+1)}$ maximizes $Q(\boldsymbol{\delta}, \boldsymbol{\delta}^{(t)})$, therefore, $Q(\boldsymbol{\delta}^{(t+1)}, \boldsymbol{\delta}^{(t)}) \geq Q(\boldsymbol{\delta}^{(t)}, \boldsymbol{\delta}^{(t)}) \Rightarrow \ell(\boldsymbol{\delta}^{(t+1)}) \geq \ell(\boldsymbol{\delta}^{(t)})$. We suggest using different starting values for the EM algorithm to check that the algorithm did not stop at a local maximum. Since the exponential

tilt parameters may be hard to interpret, it may be difficult to find initial values for them. We recommend generating an $n \times m$ matrix of initial values of the $z_{il}^{(t)}$s and starting the algorithm with the M-step.

21.4 Estimation of Features in the Component Distributions

In this section, we discuss estimation of features in the component distributions. We also identify a moment matching property similar to that found by Efron and Tibshirani (1996) in the univariate non-mixture case. For any quantity a, let $\hat{a}$ denote its estimate based on the final values of the EM algorithm at convergence. Define

$$\hat{p}_{ij} = \frac{1}{n}\left(\frac{1}{\hat{\lambda}_1 + \sum_{l=2}^{m} \hat{\lambda}_l \exp(\tilde{\mathbf{x}}_{ij}^\top \hat{\boldsymbol{\theta}}_{lj})}\right) \quad \text{and} \quad \hat{q}_{lij} = \exp(\tilde{\mathbf{x}}_{ij}^\top \hat{\boldsymbol{\theta}}_{lj})\hat{p}_{ij}.$$

The CDF of the mixture distribution, H, can be estimated by

$$\hat{H}(x_1, \ldots, x_k) = \hat{\lambda}_1 \prod_{j=1}^{k} \hat{F}_j(x_j) + \sum_{l=2}^{m} \hat{\lambda}_l \prod_{j=1}^{k} \hat{G}_{lj}(x_j),$$

where the marginal CDF estimates of F_j and G_{lj} are

$$\hat{F}_j(x_j) = \sum_{i=1}^{n} I(x_{ij} \le x_j)\hat{p}_{ij}, \quad \hat{G}_{lj}(x_j) = \sum_{i=1}^{n} I(x_{ij} \le x_j) \exp(\tilde{\mathbf{x}}_{ij}^\top \hat{\boldsymbol{\theta}}_{lj})\hat{p}_{ij}. \quad (21.22)$$

The estimates resemble the empirical CDF with the weights given by the estimated jumps. In Sect. 21.5, we give examples that show how well these estimates match the true marginal CDFs. We can also find estimates of the marginal PDFs using a weighted kernel density estimate with the posterior probabilities, $\hat{w}_{il}$, as the weights. The estimated PDFs are

$$\hat{g}_{lj}(u) = \frac{1}{\kappa} \sum_{i=1}^{n} \frac{\hat{w}_{il}}{\sum_{i=1}^{n} \hat{w}_{il}} \xi\left(\frac{u - x_{ij}}{\kappa}\right), \quad l = 2, \ldots, m \quad (21.23)$$

where κ is a bandwidth, ξ is the standard normal PDF. The R package `mixtools` contains a function, `wkde`, that allows us to do this quite easily (Young et al. 2008). This function also has the ability to choose different bandwidths for the k coordinates.

Writing $\hat{q}_{1ij} = \hat{p}_{ij}$, the mean and variance of the jth measurement in the lth component distribution, m_{lj} and s_{lj}^2 can be estimated by

$$\hat{m}_{lj} = \sum_{i=1}^{n} x_{ij}\hat{q}_{lij}, \qquad \hat{s}_{lj}^2 = \sum_{i=1}^{n} x_{ij}^2\hat{q}_{lij} - \hat{m}_{lj}^2, \tag{21.24}$$

for $l = 1, \ldots, m$ and $j = 1, \ldots, k$. An interesting result from the EM algorithm is a moment matching property. For example, we can write:

$$\begin{aligned}
\sum_{i=1}^{n} x_{ij}\hat{q}_{lij} &= \sum_{i=1}^{n} x_{ij} \exp(\tilde{\mathbf{x}}_{ij}^{\top}\hat{\boldsymbol{\theta}}_{lj}) \left\{ \frac{1}{n} \left(\frac{1}{\hat{\lambda}_1 + \sum_{l=2}^{m} \hat{\lambda}_l \exp(\tilde{\mathbf{x}}_{ij}^{\top}\hat{\boldsymbol{\theta}}_{lj})} \right) \right\} \\
&= \frac{1}{n\hat{\lambda}_l} \sum_{i=1}^{n} x_{ij} \left\{ \frac{\hat{\lambda}_l \exp(\tilde{\mathbf{x}}_{ij}^{\top}\hat{\boldsymbol{\theta}}_{lj})}{\hat{\lambda}_1 + \sum_{l=2}^{m} \hat{\lambda}_l \exp(\tilde{\mathbf{x}}_{ij}^{\top}\hat{\boldsymbol{\theta}}_{lj})} \right\} \\
&= \frac{\sum_{i=1}^{n} \hat{w}_{il} x_{ij}}{\sum_{i=1}^{n} \hat{w}_{il}},
\end{aligned}$$

where the last quantity is due to (21.18) and (21.19) from the EM algorithm. This expression matches the weighted first moment using the posterior probabilities to the tilted component first moment; see Efron and Tibshirani (1996) for an example of the moment matching property in the univariate non-mixture exponential tilt model. They argue that moment matching reduces the bias.

It should be noted that non-identifiability due to label switching (e.g., McLachlan and Peel 2000) can affect bootstrap estimation in the exponential tilt model. Suppose observations come from the following mixture

$$H(x_1, x_2, x_3) = 0.3N(0, 1)N(0, 1)N(0, 1) + 0.7N(2, 1.5)N(2.5, 2)N(3, 1).$$

Consider the first coordinate, then one possible baseline distribution is $N(0, 1)$ and the parameters in the exponential tilt would be $\boldsymbol{\theta}_{21}^{\top} = (-1.53, 1.33, 0.16)$. Another possible baseline may be $N(2, 1.5)$ in which case the parameters would be $\boldsymbol{\theta}_{21}^{T*} = (1.53, -1.33, -0.16)$. The result is that in a bootstrap, for example, the signs of the coefficients in the quadratic exponent may change. We resolve this ambiguity by designating the component corresponding to the smallest proportion as the baseline distribution and make the adjustment after the EM algorithm converges. We then have identifiable estimates of the coefficients in the quadratic exponents. The estimates of the marginal means and standard deviations are not affected by this label switching.

21.5 Simulation Results

In this section, we give simulation results for different models. The data were generated from two component mixture distributions of the following form:

$$H(x_1, x_2, x_3) = \lambda F_1(x_1)F_2(x_2)F_3(x_3) + (1-\lambda)G_1(x_1)G_2(x_2)G_3(x_3).$$

We focus on the following parameters: λ, and m_{lj} and s_{lj}, the mean and standard deviation of the jth measurement in the lth component distribution.

21.5.1 Mixtures with Normal Component Distributions

The first model is a trivariate normal mixture model, such that F_1, F_2, F_3 are CDFs of $N(0,1)$ and G_1, G_2, G_3 are CDFs of $N(\mu, \sigma^2)$ with $(\mu, \sigma^2) = (2, 1.5), (2.5, 2), (3, 1)$, respectively. Three values of $\lambda = 0.3, 0.5, 0.8$ and two different sample sizes $n = 50$ and 500 were used. For each combination of λ and n, 500 simulations were carried out. The results using $\lambda = 0.3, 0.5, 0.8$ are similar and therefore, only those under $\lambda = 0.3$ are shown. The results are given in Table 21.1, where the parameter estimates using an exponential tilt and a conditional independence normal model are given under the columns "ET" and "Normal," respectively. The exponential tilt model performs very well, its estimates are comparable to those from the normal mixture. For small samples ($n = 50$), the standard errors for the estimates using the normal model are smaller. However,

Table 21.1 Mean (standard error) of parameter estimates based on 500 simulations from a normal mixture model

		$n = 50$		$n = 500$	
	True	ET	Normal	ET	Normal
λ	0.3	0.30 (0.09)	0.31 (0.07)	0.30 (0.02)	0.30 (0.02)
m_{11}	0	0.11 (0.56)	0.02 (0.36)	0.00 (0.08)	0.00 (0.08)
m_{12}	0	0.15 (0.69)	0.01 (0.37)	−0.01 (0.09)	−0.01 (0.09)
m_{13}	0	0.22 (0.77)	0.02 (0.45)	0.02 (0.09)	0.00 (0.09)
m_{21}	2	1.97 (0.31)	1.99 (0.27)	2.00 (0.06)	1.99 (0.06)
m_{22}	2.5	2.46 (0.39)	2.49 (0.33)	2.50 (0.08)	2.49 (0.08)
m_{23}	3	2.92 (0.39)	2.98 (0.31)	2.99 (0.05)	3.00 (0.05)
s_{11}	1	0.90 (0.29)	0.92 (0.21)	0.99 (0.06)	0.99 (0.06)
s_{12}	1	0.92 (0.28)	0.93 (0.20)	0.99 (0.06)	0.99 (0.06)
s_{13}	1	1.01 (0.37)	0.95 (0.25)	1.02 (0.09)	0.99 (0.06)
s_{21}	1.22	1.18 (0.17)	1.18 (0.15)	1.21 (0.04)	1.22 (0.04)
s_{22}	1.41	1.37 (0.20)	1.37 (0.18)	1.41 (0.05)	1.41 (0.05)
s_{23}	1	1.01 (0.21)	0.97 (0.13)	0.99 (0.04)	1.00 (0.04)

the advantage of using a normal model effectively disappears for large samples ($n = 500$).

21.5.2 Mixtures with Gamma Component Distributions

Exponential tilt modeling can be thought of as a density estimation method (Efron and Tibshirani 1996). Hence we can use exponential tilt even for data that do not satisfy the exponential tilt assumption. We illustrate using mixtures of gamma distributions with different shape parameters (for application, see Dey et al. 1995; Wiper et al. 2001). We let F_1, F_2, F_3 be CDFs from a gamma(k, ζ) distribution with $(k, \zeta) = (2, 2)$, $(\mu = 4, \sigma^2 = 8)$, and G_1, G_2, G_3 are CDFs corresponding to gamma(k, ζ) distributions with $(k, \zeta) = (5, 2), (10, 1),$ and $(10, 0.5)$, respectively (with $(\mu, \sigma^2) = (10, 20), (10, 10)$, and $(5, 2.5)$, respectively). The results are also similar for different λ values, hence, we only present $\lambda = 0.4$ here. We use 1000 simulations of sample sizes $n = 50$ and 300 were carried out. We computed the estimates of the component means and standard deviations using the conditional independence normal mixture model and the conditional independence nonparametric mixture (NP) model proposed by Benaglia et al. (2009) and Levine et al. (2011) for comparison. The estimates from all three methods are shown in Table 21.2. When the sample size is small, the performance of the exponential tilt method is similar to the normal mixture model. For larger sample size ($n = 300$), the tilted method does much better than the normal model and follows more closely to the nonparametric method.

Table 21.2 Mean (standard error) of parameter estimates based on 1000 simulations from a gamma mixture model

		$n = 50$			$n = 300$		
	True	ET	Normal	NP	ET	Normal	NP
λ	0.4	0.39 (0.12)	0.37 (0.13)	0.38 (0.10)	0.37 (0.04)	0.32 (0.05)	0.36 (0.04)
m_{11}	4	3.86 (1.02)	3.66 (0.99)	3.86 (0.85)	3.85 (0.35)	3.41 (0.36)	3.78 (0.31)
m_{12}	4	4.78 (2.52)	4.53 (2.68)	4.47 (2.15)	3.78 (0.53)	3.32 (0.45)	3.71 (0.34)
m_{13}	4	4.03 (0.89)	4.11 (0.93)	4.02 (0.79)	3.97 (0.30)	4.06 (0.35)	3.96 (0.29)
m_{21}	10	9.99 (1.12)	10.03 (1.38)	9.93 (1.19)	9.84 (0.38)	9.56 (0.39)	9.76 (0.36)
m_{22}	10	9.04 (1.77)	9.04 (1.66)	9.27 (1.61)	9.86 (0.41)	9.59 (0.32)	9.79 (0.28)
m_{23}	5	4.84 (0.44)	4.80 (0.42)	4.86 (0.42)	4.96 (0.13)	4.84 (0.13)	4.95 (0.12)
s_{11}	2.82	2.47 (1.05)	2.11 (0.86)	2.63 (0.96)	2.65 (0.44)	1.96 (0.28)	2.62 (0.40)
s_{12}	2.82	2.47 (1.00)	2.11 (0.87)	2.61 (0.91)	2.53 (0.50)	1.90 (0.33)	2.54 (0.38)
s_{13}	2.82	2.42 (0.90)	2.40 (0.95)	2.49 (0.79)	2.84 (0.32)	2.91 (0.39)	2.80 (0.32)
s_{21}	4.47	4.23 (0.80)	4.23 (0.87)	4.31 (0.72)	4.48 (0.29)	4.58 (0.28)	4.50 (0.28)
s_{22}	3.16	3.20 (0.62)	3.27 (0.58)	3.15 (0.51)	3.21 (0.22)	3.42 (0.23)	3.27 (0.20)
s_{23}	1.58	1.69 (0.46)	1.75 (0.43)	1.70 (0.35)	1.60 (0.14)	1.71 (0.13)	1.67 (0.12)

Table 21.3 Mean (standard error) of parameter estimates based on 1000 simulations from a mixture model of normal and gamma

		$n = 50$			$n = 300$		
	True	ET	Normal	NP	ET	Normal	NP
λ	0.4	0.33(0.08)	0.34(0.11)	0.36(0.09)	0.30 (0.03)	0.35 (0.04)	0.35 (0.04)
m_{11}	0	0.91(1.61)	1.57(2.04)	1.50(1.90)	0.04 (0.16)	0.27 (0.23)	0.39 (0.22)
m_{12}	4	3.23(1.60)	2.55(1.79)	2.70(1.74)	3.94 (0.31)	3.45 (0.41)	3.64 (0.33)
m_{13}	0	−0.04(1.48)	−0.15(3.69)	−0.07(3.37)	0.00 (0.11)	−0.01 (0.15)	0.00 (0.23)
m_{21}	4	3.69(0.98)	3.49(1.23)	3.52(1.22)	4.00 (0.20)	4.13 (0.23)	4.04 (0.20)
m_{22}	0	0.23(0.86)	0.48(0.93)	0.39(1.05)	0.01 (0.10)	0.02 (0.11)	−0.07 (0.10)
m_{23}	0	−0.01(1.04)	−0.01(0.98)	−0.01(1.06)	0.00 (0.38)	0.01 (0.40)	0.01 (0.42)
s_{11}	1	1.61(1.07)	1.63(1.02)	1.97(1.00)	1.15 (0.41)	1.12 (0.18)	1.75 (0.48)
s_{12}	2.82	2.45(0.91)	2.43(1.19)	2.40(0.91)	2.78 (0.33)	3.03 (0.34)	2.82 (0.31)
s_{13}	1	1.85(1.85)	2.54(2.53)	2.88(2.12)	0.99 (0.11)	1.07 (0.56)	2.06 (0.81)
s_{21}	2.82	2.61(0.62)	2.62(0.69)	2.56(0.63)	2.79 (0.23)	2.87 (0.23)	2.75 (0.22)
s_{22}	1.41	1.46(0.59)	1.61(0.76)	1.55(0.69)	1.42 (0.15)	1.33 (0.12)	1.33 (0.13)
s_{23}	5.65	5.10(1.37)	4.69(1.63)	4.76(1.43)	5.64 (0.44)	5.79 (0.52)	5.60 (0.44)

21.5.3 Mixtures with Different Component Distributions

The third set of simulations studied the situation where the marginals are from different families of distributions (see, e.g., Khalili et al. 2007). We let F_1, F_2, F_3 be CDFs from N(0,1), gamma($k = 2, \zeta = 2$), ($\mu = 4, \sigma^2 = 8$), and N(0,1) distributions, respectively, and G_1, G_2, G_3 are CDFs corresponding to gamma($k = 2, \zeta = 2$), Laplace distributions with location and scale parameters (0,1), ($\mu = 0, \sigma^2 = 2$), and a Laplace with parameters (0,4), ($\mu = 0, \sigma^2 = 32$), respectively. The results under different values of λ are similar and hence only results for $\lambda = 0.3$ are given. One thousand simulations of sample sizes $n = 50$ and 300 were carried out. The results are given in Table 21.3.

It can be observed that the tilted method produces the best results for nearly all the parameters. We also plotted the estimated marginal CDFs and PDFs for one of the simulations in Fig. 21.1.

Again, even though the exponential tilt assumption does not hold here, the exponential tilt estimates of the component means, standard deviations, CDF, and PDF are very good.

21.6 Model Selection

In this section, we show how to estimate the number of components in the mixture. We use a modified BIC (Bayesian Information Criterion, Schwarz 1978) model selection criterion pBIC $\equiv -2 \ln L_p + s \ln(n)$, where L_p is the maximized

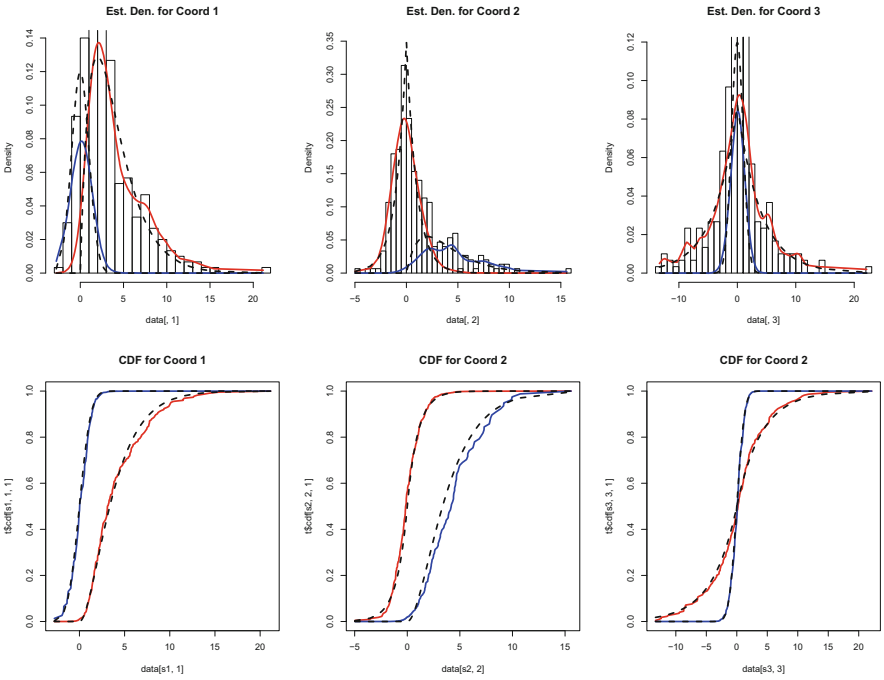

Fig. 21.1 Semiparametric estimation for a randomly selected dataset from the simulations with different distributions with $n = 300$. The *dotted line* represents the true CDFs and PDFs. In this dataset, the estimates are: $\hat{\lambda} = 0.25$, $(\hat{m}_{11}, \hat{m}_{12}, \hat{m}_{13}) = (0.10, 4.49, 0.06)$, $(\hat{m}_{21}, \hat{m}_{22}, \hat{m}_{23}) = (4.20, -0.04, 0.14)$, $(\hat{s}_{11}, \hat{s}_{12}, \hat{s}_{13}) = (1.08, 2.86, 1.02)$, $(\hat{s}_{21}, \hat{s}_{22}, \hat{s}_{23}) = (3.28, 1.37, 5.05)$

semiparametric profile likelihood and s is the number of parameters in the model. Since mixture models do not satisfy all the regularity conditions in Schwarz (1978) we turn to simulations to study the criterion. We use three models for simulations:

Model 1: Normal location mixtures with $m = 2, 3, 4$ components. There are $k = 7$ repeated measures with $(m_{1j}, m_{2j}, m_{3j}, m_{4j}) = (0, 2, 4, 6)$ and $s_{lj} = 1$ for $l = 1, \ldots, m; j = 1, \ldots, 7$.

Model 2: Normal location mixtures with $m = 2, 3, 4$ components. There are $k = 10$ repeated measures with $(m_{1j}, m_{2j}, m_{3j}, m_{4j}) = (0, 2, 4, 6)$ and $s_{lj} = 1$ for $l = 1, \ldots, m; j = 1, \ldots, 10$.

Model 3: Normal scale mixtures with $m = 2, 3$ components. There are $k = 5$ repeated measures with $(m_{1j}, m_{2j}, m_{3j}) = (0, 0, 0)$ and $(s_{1j}, s_{2j}, s_{3j}) = (0, 10, 50)$ for $j = 1, \ldots, 5$.

Table 21.4 gives the proportion of times pBIC selected the correct number of components. For each model considered, the mixing proportions of the components are equal, i.e., for a model with m components, $\lambda_1 = \lambda_2 = \ldots = \lambda_m$. Included in the table is the number of parameters estimated in each model, $N_p = 3k(m-1)+(m-1)$, which includes the exponential tilt parameters for each of k dimensions in the $m-1$

Table 21.4 pBIC simulations results for Models 1–3 where n is the number of observations, k is the number of repeated measures, and m is the true number of components

		$m = 2$			$m = 3$			$m = 4$		
Model	k	N_p	$n = 100$	$n = 200$	N_p	$n = 100$	$n = 200$	N_p	$n = 100$	$n = 200$
1	7	15	1.00	1.00	30	1.00	1.00	45	0.96	0.98
2	10	21	1.00	1.00	42	0.95	0.99	63	0.91	0.97
3	5	11	0.94	0.97	22	0.65	0.67	—	—	—

components and the $m - 1$ mixing proportions λ_l. For Model 3 with $m = 3$ the success rate for pBIC was roughly $2/3$. However, when the sample size increased to 500, the success rate increased to 0.90. As a check, we compared pBIC to a modified Akakie Information Criterion, which gives similar results. We conclude that pBIC is effective for estimating the number of components in the semiparametric mixture.

21.7 Example

We applied the proposed method to a real data problem. The data comes from a cognitive experiment discussed in Cruz-Medina et al. (2004) and is available at http://www.blackwellpublishing.com/rss. The experiment was used to demonstrate children fall into different groups in their approach to solve cognitive tasks. The experiment recruited normally developing 9-year-old children. Each child was given a set of different task conditions, which is a visual stimulus that involves two images on a computer monitor. The left image is the target stimulus and the right image is either identical to the target image or the mirror image of the target stimulus. The child pressed one key indicating if he/she thought the right image was identical or another key if they thought it was the mirror image. The outcome of interest is the reaction time (RT), in milliseconds, for a child to give a response to the visual stimulus. Each child was given $k = 6$ different task conditions and the RT for the child to choose the correct response on each task was recorded. We focused on the subset of $n = 197$ children who gave correct responses to all the task conditions. Since the six task conditions were embedded in a random sequence of tasks, the children could not have anticipated which task condition would appear. Therefore, given that a child was in a particular group, it would not be unreasonable to assume that their reaction times were independent and the conditional independence assumption seems valid. Longer response times may indicate reading comprehension problems. See Miller et al. (2001) for additional background on this experiment.

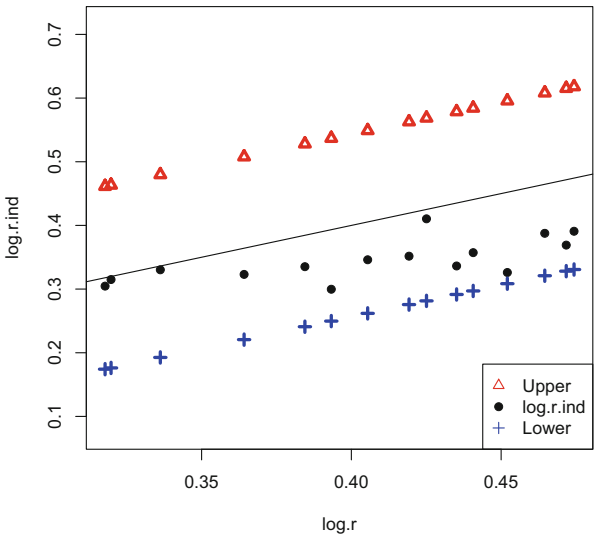

Fig. 21.2 Plot of the transformed sample correlations (log.r) against the transformed sample correlations calculated under conditional independence (log.r.ind). The *upper* and *lower* confidence bands of the transformed sample correlations are shown. The *solid line* is the 45° line

To further examine whether the conditional independence assumption is reasonable, we looked at the sample correlations between the coordinates and those under conditional independence. The correlations were calculated and Fisher's transformations were performed. Figure 21.2 plots the transformed correlations under conditional independence against the transformed correlations with no assumptions.

As a rough check, we included upper and lower points computed using $2/\sqrt{n-3}$ as an estimate of the standard errors of the transformed correlations. All estimates assuming conditional independence fall within these bounds.

We compared $m = 1, 2, 3, 4$ component models for this dataset using pBIC and selected $m = 3$ based on its lowest pBIC value ($\ell_P = -6081.6$, pBIC$= 12300.6$ and corresponding number of parameters $= 26$). The data based on $m = 3$ can be written as $x_{ij}, i = 1, 2, \ldots, 197; j = 1, \ldots, 6$ with corresponding CDF

$$H(x_1, \ldots, x_6) = \lambda_1 \prod_{j=1}^{6} F_j(x_j) + \sum_{l=2}^{3} \lambda_l \prod_{j=1}^{6} G_{lj}(x_j).$$

The estimated marginal CDFs, means and standard deviations of $F_j, G_{2j}, G_{3j}, j = 1, \ldots, 6$ using (21.22) and (21.24) are given in Table 21.5.

It appears that the distribution of RTs for the first task condition may well be different from the distributions of RTs for the other task conditions. From the results, the smallest group of children composed of about 20 % appear to have the shortest RTs and also the smallest variation. This might suggest that these children

Table 21.5 Estimated component means for the RT data with $m = 3$ components

Component					
1		2		3	
λ_1	0.49	λ_2	0.20	λ_3	0.31
m_{11}	2024	m_{21}	1577	m_{31}	3024
m_{12}	1712	m_{22}	1456	m_{32}	2776
m_{13}	1864	m_{23}	1265	m_{33}	2761
m_{14}	1799	m_{24}	1312	m_{34}	2771
m_{15}	1870	m_{25}	1171	m_{35}	2729
m_{16}	1957	m_{26}	1216	m_{36}	2661
s_{11}	691	s_{21}	420	s_{31}	1074
s_{12}	469	s_{22}	337	s_{32}	907
s_{13}	609	s_{23}	200	s_{33}	1101
s_{14}	516	s_{24}	332	s_{34}	1097
s_{15}	777	s_{25}	402	s_{35}	1162
s_{16}	636	s_{26}	261	s_{36}	1180

understand the concept and are quick to choose correctly. The next group composed of about 30 % of the children have the longest RTs as well as the largest variation. For the children in this group, a possible explanation is that they look longer to react to certain tasks and quicker for other tasks. It would be interesting to break up the trials based on which was the correct answer, the identical image or the mirror image. The last and largest group, about 50 %, are the children in the middle.

Figure 21.3 shows the semiparametric estimates of the component CDFs. Similar analyses were carried out using the log transformed data with similar results. Note the variation in the coordinate means and standard deviations again suggests that the component marginal distributions differ. The data were originally analyzed by Cruz-Medina et al. (2004) by discretizing the data and assuming that the repeated measures were identically distributed. The estimated proportions were 0.55, 0.16, and 0.29 in the order given in Table 7. The common coordinate medians were 1689, 1273, and 2523 for the three components and are a bit lower than the reported sets of five means for each component.

21.8 Discussion and Modifications

Walther (2002) introduces a univariate mixture of log-concave densities. He gives a representation theorem, and based on this theorem, develops a test for the presence of a mixture model. Chang and Walther (2007) extend this model to the multivariate case in (21.1). However, lack of identifiability is a difficulty for their model (Walther 2002, pp. 509); a mixture of log-concave densities may itself be log-concave and identifiability fails.

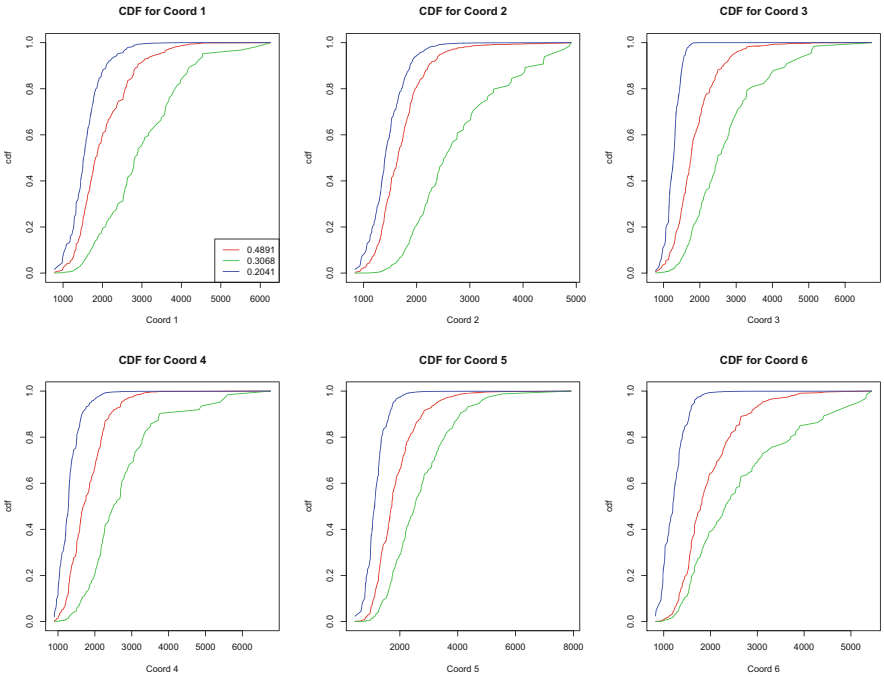

Fig. 21.3 Semiparametric estimates of $F_j, G_{lj},\ j = 1, \ldots, 6,\ l = 2, 3$ under the exponential tilt model for the RT data

The method we proposed assumes all repeated measures are related by (21.3). Our model could be modified to handle situations where some of the dimensions are modeled by (21.3) while the others are modeled parametrically. For example, let the first j_1 dimensions be modeled by (21.3), then

$$
\begin{aligned}
&h(x_1, \ldots, x_k, \boldsymbol{\delta}, \boldsymbol{\Omega}) \\
&= \left\{ \lambda_1 \prod_{j=j_1+1}^{k} f_j(x_j, \omega_j^f) + \sum_{l=2}^{m} \lambda_l \exp\Big(\sum_{j=1}^{j_1} \tilde{\mathbf{x}}_j^\top \boldsymbol{\theta}_{lj}\Big) \prod_{j=j_1+1}^{k} g_j(x_j, \omega_j^g) \right\} \prod_{j=1}^{j_1} f_j(x_j) \\
&= \psi(x_1, \ldots, x_k, \boldsymbol{\delta}, \boldsymbol{\Omega}) \prod_{j=1}^{j_1} f_j(x_j),
\end{aligned}
$$

where $f_j, g_j, j = j_1 + 1, \ldots, k$ are parametrized by $\boldsymbol{\Omega} = (\omega_j^f, \omega_j^g)$, which leads to

$$
\ell_p(\boldsymbol{\delta}, \boldsymbol{\Omega}) = \log \psi(x_1, \ldots, x_k, \boldsymbol{\delta}, \boldsymbol{\Omega}) - \sum_{i=1}^{n} \sum_{j=1}^{j_1} \log \left[n + \sum_{l=2}^{m} \eta_{lj} \{ \exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1 \} \right].
$$

Our formulation of a multivariate mixture can also be interpreted as a copula. Replacing η_{lj}/n by λ_l, a simple rearrangement yields the profile likelihood as

$$n^{-kn} \prod_{i=1}^{n} \left[\frac{\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj})}{\prod_{j=1}^{k} \{\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj})\}} \right].$$

If we multiply the numerator and denominator by $\prod_{j=1}^{k} f_j(x_j)$, then the profile likelihood is proportional to $c\,(H_1(x_1), \ldots, H_k(x_k)) \equiv h(x_1, \ldots, x_k) / \prod_{j=1}^{k} h_j(x_j)$, the joint mixture density divided by a product of the marginal densities. This can be viewed as a semiparametric copula density evaluated at the marginal CDFs. We can also interpret our exponential tilt mixture as $h(x_1, \ldots, x_k) = c\,(H_1(x_1), \ldots, H_k(x_k)) \prod_{j=1}^{k} h_j(x_j)$. Hence we begin with a product of (independent) marginals and model the correlation and mixture structure via the copula based on the mixture of exponential tilts. Further motivation can be found in Chen et al. (2006). This approach also avoids the curse of dimensionality problem associated with estimation in high dimensional distributions.

Acknowledgements Tracey Wrobel Hammel and Thomas Hettmansperger were partially supported by NSF Grant SES-0518772. Denis Leung was supported by SMU Research Center.

References

Allman, E.S., Matias, C., Rhodes, J.A.: Identifiability of parameters in latent class models with many observed variables. Ann. Stat. **37**, 3099–3132 (2009)

Anderson, J.A.: Multivariate logistic compounds. Biometrika **66**, 17–26 (1979)

Benaglia, T., Chauveau, D., Hunter, D.R.: An EM-like algorithm for semi- and non-parametric estimation in multivariate mixtures. J. Comput. Graph. Stat. **18**, 505–526 (2009)

Chang, G.T., Walther, G.: Clustering with mixtures of log-concave distributions. Comput. Stat. Data Anal. **51**, 6242–6251 (2007)

Chen, X., Fan, Y., Tsyrennikov, V.: Efficient estimation of semiparametric multivariate copula models. J. Am. Stat. Assoc. **101**, 1228–1240 (2006)

Cruz-Medina, I.R., Hettmansperger, T.P., Thomas, H.: Semiparametric mixture models and repeated measures: the multinomial cut point model. Appl. Stat. **53**, 463–474 (2004)

Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B **39**, 1–38 (1977)

Dey, D., Kuo, L., Sahu, S.: A bayesian predictive approach to determining the number of components in a mixture distribution. Stat. Comput. **5**, 297–305 (1995)

Efron, B., Tibshirani, R.: Using specially designed exponential families for density estimation. Ann. Stat. **24**, 2431–2461 (1996)

Elmore, R.T., Hall, P., Neeman, A.: An application of classical invariant theory to identifiability in nonparametric mixtures. Ann. I. Fourier **55**, 1–28 (2005)

Hall, P., Neeman, A., Pakyari, R., Elmore, R.T.: Nonparametric inference in multivariate mixtures. Biometrika **92**, 667–678 (2005)

Hall, P., Zhou, X.H.: Nonparametric estimation of component distributions in a multivariate mixture. Ann. Stat. **31**, 201–224 (2003)

Hettmansperger, T.P., Thomas, H.: Almost nonparametric inference for repeated measure in mixture models. J. R. Stat. Soc. Ser. B **62**, 811–825 (2000)
Kay, R., Little, S.: Assessing the fit of the logistic model: A case study of children with the haemolytic uraemic syndrome. Appl. Stat. **35**, 16–30 (1986)
Khalili, A., Potter, D., Yan, P., Li, L., Gray, J., Huang, T., Lin, S.: Gamma-normal-gamma mixture model for detecting differentially methylated loci in three breast cancer cell lines. Cancer Informat. **3**, 43–54 (2007)
Leung, D., Qin, J.: Semi-parametric inference in a bivariate (multivariate) mixture model. Stat. Sin. **16**, 153–163 (2006)
Levine, M., Hunter, D.R., Chauveau, D.: Maximum smoothed likelihood for multivariate mixtures. Biometrika **98** 403–416 (2011)
Lindsay, B.G.: Mixture Models: Theory, Geometry, and Applications. Institute of Mathematical Statistics, Hayward, CA (1995)
Liu, Q., Pierce, D.A.: A note on gauss-hermite quadrature. Biometrika **81**, 624–629 (1994)
McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
Miller, C.A., Kail, R., Leonard, L.B., Tomblin, J.B.: Speed of processing in children with specific language impairment. J. Speech Lang. Hear. Res. **44**, 416–433 (2001)
Nagelkerke, N.J.D., Borgdorff, M.W., Kim, S.J.: Logistic discrimination of mixtures of m. tuberculosis and non-specific tuberculin reactions. Stat. Med. **20**, 1113–1124 (2001)
Owen, A.: Empirical likelihood ratio confidence intervals for a single functional. Biometrika **75**, 237–49 (1988)
Qin, J., Berwick, M., Ashbolt, R., Dwyer, T.: Quantifying the change of melanoma incidence by breslow thickness. Biometrics **58**, 665–670 (2002)
Qin, J., Zhang, B.: A goodness-of-fit test for logistic regression models based on case-control data. Biometrika **84**, 609–618 (1997)
Qu, Y.S., Hadgu, A.: A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. J. Am. Stat. Assoc. **93**, 920–928 (1998)
Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **5**, 461–464 (1978)
Thomas, H., Lohaus, A.: Modeling growth and individual differences in spatial tasks. Monogr. Soc. Res. Child Dev. **58**, 1–191 (1993)
Titterington, D.M., Smith, A.F.M., Makov, U.E.: Statistical Analysis of Finite Mixture Distributions. Wiley, Chichester (1985)
Walther, G.: Detecting the presence of mixing with multiscale maximum likelihood. J. Am. Stat. Assoc. **97**, 508–513 (2002)
White, I.R., Thompson, S.G.: Choice of test for comparing two groups, with particular application to skewed outcomes. Stat. Med.**21**, 1205–1215 (2003)
Wiper, M., Rios Insua, D., Ruggeri, F.: Mixtures of gamma distributions with applications. J. Comput. Graph. Stat. **10**, 440–454 (2001)
Young, D.S., Benaglia, T., Chauveau, D., Elmore, R.T., Hettmansperger, T.P., Hunter, D.R., Thomas, H., Xuan, F.: mixtools: Tools for mixture models. R package version 0.3.2 (2008)
Zhang, B.: An information matrix test for logistic regression models based on case-control data. Biometrika **88**, 921–932 (2001)

Part IV
Invariant Coordinate Selection and Related Methods

Chapter 22
A B-Robust Non-Iterative Scatter Matrix Estimator: Asymptotics and Application to Cluster Detection Using Invariant Coordinate Selection

Mohamed Fekri and Anne Ruiz-Gazen

Abstract In Ruiz-Gazen (Comput Stat Data Anal 21:149–162, 1996), a simple B-robust estimator was introduced. Its definition is explicit and takes into account the empirical covariance matrix together with a one-step M-estimator. In the present paper, we derive the asymptotics and some robustness properties of this estimator. We compare its performance to the usual M- and S-estimators by means of a Monte-Carlo study. We also illustrate its use for cluster detection using Invariant Coordinate Selection on a small example.

Keywords Asymptotic distribution • Eigenelements • Influence function • M-estimators • MCD estimators • S-estimators

22.1 Introduction

The empirical mean and covariance estimators are the maximum likelihood estimators of multivariate gaussian distribution parameters. But it is well known that these estimators are very sensitive to outlying observations (Devlin et al. 1981; Rousseeuw and Leroy 1987). Hence, many authors have proposed robust location and scatter estimators (see Maronna and Yohai 1988 for a survey). The most widely used robustness criteria are the B-robustness (bounded influence function) and a high breakdown point. While, roughly speaking, the influence function is a local robustness measure which describes the effect of infinitesimal contamination on

M. Fekri
Département de Mathématiques, Informatique et Réseaux, Institut National des Postes et Télécommunications, Avenue Allal Al Fassi, Rabat, Maroc
e-mail: fekri@menara.ma

A. Ruiz-Gazen (✉)
Toulouse School of Economics, Université Toulouse 1 Capitole, 21, allée de Brienne, 31000 Toulouse, France
e-mail: anne.ruiz-gazen@tse-fr.eu

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_22

the estimator, the breakdown point is a global notion which gives the smallest proportion of outliers that can carry the estimator over all bounds.

Robust location and scatter estimators may be classified in different categories. The first category regroups the "element-wise" estimators which consist in estimating robustly each variance and covariance term and putting them altogether in order to reconstitute a dispersion matrix estimator. One of the first proposals in this category is studied in Devlin et al. (1981) and other references are Ma and Genton (2001), Maronna and Zamar (2002), Béguin and Hulliger (2004). The main advantage of the element-wise approach is its low computational cost. However, its drawback is that the obtained dispersion estimators are not necessarily definite positive nor affine equivariant.

The second category consists in affine equivariant estimators which are defined in a global way. Among them are the M-, the S-, the Minimum Volume Ellipsoid (MVE), and the Minimum Covariance Determinant (MCD) estimators, their reweighted version and the projection based estimators. Multivariate location and dispersion M-estimators generalize the maximum likelihood location and dispersion estimators of an elliptical model. They are defined as weighted means and covariances where the weights are decreasing functions of the Mahalanobis distances based on the robust estimators themselves. So, their defining equations are implicit and optimization algorithms are needed in order to calculate them. Their asymptotic and robustness properties have been studied notably in Maronna (1976). They are B-robust estimators but their breakdown point is decreasing when the dimension increases. Actually, if p denotes the dimension of the data, the breakdown point of M-estimators is $1/p$. These estimators have been revisited recently in Dümbgen et al. (2013a), Dümbgen et al. (2013b).

On the contrary, the S-, MVE, and MCD estimators can reach a high breakdown point even in high dimensions. These estimators are also B-robust. The MVE and the MCD estimators have been introduced in Rousseeuw (1985), Rousseeuw and Leroy (1987). Both are high breakdown location and dispersion estimators but, since the MVE is not $\sqrt{n}$ convergent (see Davies 1992), it has been more or less superseded by the MCD estimator which has the usual rate of convergence (Cator and Lopuhaä 2012). The MVE and the MCD estimators have been generalized to the class of S-estimators by Davies (1987). The counterpart of using high breakdown point estimators is the lack of efficiency under the assumed model (generally the gaussian model). In order to increase the efficiency of high breakdown point estimators, some authors have proposed to use reweighted estimators (see Rousseeuw and Croux 1994). The idea is to use high breakdown point estimators as an initial step and calculate weighted location and dispersion estimators. Usually, the weight is a dummy variable which equals zero for observations associated with large Mahalanobis distances. The asymptotic and the robustness properties of reweighted estimators have been studied in Lopuhaä (1999). Croux and Haesbroeck (2000) give an interesting comparison of the efficiency of the reweighted S- and MCD estimators and conclude that reweighting the MCD estimator leads to an important gain in efficiency while it is not really worthwhile for S-estimators (except for small dimension with 25 % breakdown point).

The main drawback of S- and MCD estimators is that they are based on objective functions that may have many local optima. Consequently, they are calculated through approximate algorithms which use random subsampling (Rousseeuw and Leroy 1987). Multivariate dispersion matrix estimators, called sign and rank covariance matrix estimators, with high efficiency at elliptical models have been introduced in Visuri et al. (2000) and studied in Ollila et al. (2002), Ollila et al. (2003). They have not only a zero breakdown point but also an unbounded influence function.

The projection based estimators constitute another class of estimators which are based on the idea that, if a multivariate observation is outlying, there must exist a one-dimensional projection of the data on which this observation is a univariate outlier. The first estimator defined in this category is the Stahel–Donoho estimator proposed independently by Stahel (1981), Donoho (1982). Maronna and Yohai (1995) have proved that they are $\sqrt{n}$-consistent, studied their asymptotic bias and illustrated via simulation that they have a high relative efficiency. Gervini (2002) derived their influence functions and noted that they have the same form as influence functions of one-step estimators which may explain their high relative efficiency. The asymptotic normality of the Stahel–Donoho estimator has been proved by Zuo et al. (2004).

From a computational point of view, the Stahel–Donoho estimator, as many projection-pursuit based estimators, is difficult to calculate since all the projection directions have to be considered. Zuo and Lai (2011) is a recent reference on the subject but only for bivariate data. Peña and Prieto (2001) proposed to consider a certain set of projection directions so that "the proposed procedure can be seen as an empirically successful and faster way of implementing the Stahel–Donoho algorithm."

Note also that another class of projection based estimators (called P-estimators) has been introduced by Maronna et al. (1992) but they have not been studied further since their numerical computation involves a double optimization procedure which is very difficult to handle.

As a summary on robust multivariate location and dispersion estimation, we can say that high breakdown point estimators are attractive but are difficult to compute numerically.

In Ruiz-Gazen (1996), a simple scatter estimator denoted by $\mathbf{U}_n(\beta)$ with $\beta > 0$ is introduced. It consists of a simple transformation of a one-step weighted dispersion matrix. The main advantage of this estimator is that its definition is explicit and it is very easy and fast to compute. It has a zero breakdown point but is B-robust. In the present paper, we propose to complete the asymptotic study of this estimator and provide an application to cluster detection using Invariant Coordinate Selection (ICS) as proposed in Caussinus et al. (2003) and generalized in Tyler et al. (2009).

The paper is organized as follows. In the second section, we recall the definition of $\mathbf{U}_n(\beta)$ and prove its consistency in the gaussian case. We also prove its asymptotic normality when the underlying distribution has fourth moments and give an explicit asymptotic variance in the gaussian case. We also give the asymptotic distribution of the eigenelements of $\mathbf{U}_n(\beta)$ even in the case of equal eigenvalues which is not

standard. In a third section, we use the perturbation theory (Kato 1966) to calculate the influence function of $\mathbf{U}_n(\beta)$ and of its eigenelements. Section 22.4 contains the results of a Monte Carlo study which shows that $\mathbf{U}_n(\beta)$ is calculated much faster than the S-estimators and is competitive when its results are compared to M-estimators. Section 22.5 applies $\mathbf{U}_n(\beta)$ on a small data set in order to detect clusters using ICS. In this example the use of high breakdown point scatter estimators is not adapted while there is a gain in using the proposed estimator compared to the empirical covariance estimator.

22.2 Definition, Consistency, and Asymptotic Normality

22.2.1 Definition

Let us consider $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, n observations in $\mathbf{R}^p$ ($n > p$) independent and identically distributed. The common distribution F has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ assumed to be positive definite. For $\beta > 0$, we define the covariance matrix estimator $\mathbf{U}_n(\beta) = \left(\mathbf{S}_n(\beta)^{-1} - \beta \mathbf{V}_n^{-1}\right)^{-1}$ with:

$$\mathbf{V}_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu}_n)(\mathbf{X}_i - \boldsymbol{\mu}_n)^\top,$$

$$\mathbf{S}_n(\beta) = \frac{\sum_{i=1}^{n} K(\beta \|\mathbf{X}_i - \boldsymbol{\mu}_n\|^2_{\mathbf{V}_n^{-1}})(\mathbf{X}_i - \boldsymbol{\mu}_n)(\mathbf{X}_i - \boldsymbol{\mu}_n)^\top}{\sum_{i=1}^{n} K(\beta \|\mathbf{X}_i - \boldsymbol{\mu}_n\|^2_{\mathbf{V}_n^{-1}})},$$

$$\|\mathbf{X}_i - \boldsymbol{\mu}_n\|^2_{\mathbf{V}_n^{-1}} = (\mathbf{X}_i - \boldsymbol{\mu}_n)^\top \mathbf{V}_n^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_n),$$

K a decreasing function from $\mathbf{R}^+$ to $\mathbf{R}^+$ and $\boldsymbol{\mu}_n$ a location estimator which converges almost surely (a.s.) to $\boldsymbol{\mu}$.

We recall from Ruiz-Gazen (1996) that:

1. if $\boldsymbol{\mu}_n$ is an affine equivariant location estimator, then $\mathbf{U}_n(\beta)$ is an affine equivariant covariance matrix estimator,
2. the definition of $\mathbf{U}_n(\beta)$ may be justified by studying the almost sure limit of $\mathbf{U}_n(\beta)$ to $\boldsymbol{\Sigma}$ for a contaminated model of the form $(1-\varepsilon)\mathcal{N}(\mathbf{x}_0, \boldsymbol{\Sigma}) + \varepsilon\mathcal{N}(\mathbf{x}_1, \boldsymbol{\Sigma})$, where $\mathbf{x}_0$ and $\mathbf{x}_1$ are vectors in $\mathbf{R}^p$, $\varepsilon \in]0;1[$ and $K(x) = \exp(-x/2)$. Moreover, for any $\beta > 0$, the influence function of $\mathbf{U}_n(\beta)$ is bounded while the influence function of $\mathbf{S}_n(\beta)$ is not.
3. $\mathbf{U}_n(\beta)$ is positive definite for $\beta \leq 1/p$. So, we may use $\beta = 1/p$ for small to moderate values of p. A Monte Carlo study has shown that the use of $\beta_0 = (26\varepsilon)/p$, where ε is the proportion of contaminated data, may be advisable for moderate to large values of p, even if β_0 is more than $1/p$. In the present paper, we propose to use $\varepsilon = 5\,\%$ or $\varepsilon = 10\,\%$ but this percentage may be increased if one suspects that the data are heavily contaminated.

In this paper, we use the following notations:

- $\text{vec}\boldsymbol{\Sigma}$ denotes the column vector obtained by piling up the columns of the matrix $\boldsymbol{\Sigma}$,
- $\otimes$ denotes the Kronecker product,
- $\mathbf{C}_p$ denotes the commutation matrix $p^2 \times p^2$ such that $\mathbf{C}_p \times \text{vec}\mathbf{A} = \text{vec}\mathbf{A}^\top$.

In the following, we obtain the almost sure convergence and the asymptotic normality of $\mathbf{U}_n(\beta)$ under weak hypotheses on the weight function K and on the distribution F. In order to obtain consistency and explicit asymptotic variances, we focus on the gaussian distribution and use an exponential form for the weight function. The use of such a weight function has been suggested by Meshalkin (1970) and leads to simple expressions in the gaussian case.

Proofs of Theorems 22.1 and 22.2 rely on the following lemma by Le Cam (1953) which is a version of the uniform law of large numbers.

Lemma 22.1 *Let $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots,$ be independent and identically distributed random variables and $g(\mathbf{x}, \boldsymbol{\theta})$ a continuous function in $\boldsymbol{\theta} \in \Theta$ where Θ is a compact set. If there exists a function H such that $E(H(\mathbf{X})) < \infty$ and $\|g(\mathbf{x}, \boldsymbol{\theta})\| \leq H(\mathbf{x})$, for all $\mathbf{x}$ and $\boldsymbol{\theta}$, then $\sup_{\boldsymbol{\theta} \in \Theta} \| \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{X}_i, \boldsymbol{\theta}) - E(g(\mathbf{X}, \boldsymbol{\theta})) \|$ converges a.s. to 0.*

22.2.2 Consistency

Theorem 22.1 gives the almost sure limit of $\mathbf{S}_n(\beta)$ and $\mathbf{U}_n(\beta)$ for general weight functions K.

Theorem 22.1 *Let us assume that the distribution F of $\boldsymbol{X}$ has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. If K is a bounded function on $\mathbf{R}^+$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mapsto K(\beta \|\mathbf{X} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})$ is almost surely continuous for all $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{S}_n(\beta)$ converges a.s. to*

$$\mathbf{S}(\beta) = \frac{E_F\left(K(\beta \|\mathbf{X} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top\right)}{E_F\left(K(\beta \|\mathbf{X} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})\right)} \tag{22.1}$$

and $\mathbf{U}_n(\beta)$ converges a.s. to $(\mathbf{S}^{-1}(\beta) - \beta \boldsymbol{\Sigma}^{-1})^{-1}$.

Proof

$$\text{Let } \mathbf{A}_n(\boldsymbol{\mu}_n, \mathbf{V}_n) = \frac{1}{n} \sum_{i=1}^{n} K(\beta \|\mathbf{X}_i - \boldsymbol{\mu}_n\|^2_{\mathbf{V}_n^{-1}})(\mathbf{X}_i - \boldsymbol{\mu}_n)(\mathbf{X}_i - \boldsymbol{\mu}_n)^\top,$$

$$\text{and } a_n(\boldsymbol{\mu}_n, \mathbf{V}_n) = \frac{1}{n} \sum_{i=1}^{n} K(\beta \|\mathbf{X}_i - \boldsymbol{\mu}_n\|^2_{\mathbf{V}_n^{-1}}).$$

We have

$$\mathbf{S}_n(\beta) = \frac{\mathbf{A}_n(\boldsymbol{\mu}_n, \mathbf{V}_n)}{a_n(\boldsymbol{\mu}_n, \mathbf{V}_n)}.$$

Let $\boldsymbol{\theta} = (\boldsymbol{\mu}^\top, (\mathrm{vec}\boldsymbol{\Sigma})^\top)^\top$. We denote by $B_r(\boldsymbol{\theta})$ the closed ball of center $\boldsymbol{\theta}$ and radius r. As $\boldsymbol{\mu}_n$ converge a.s. to $\boldsymbol{\mu}$, for $r > 0$, with probability one, there is an integer N such that for $n > N$, $\boldsymbol{\theta}_n = (\boldsymbol{\mu}_n^\top, (\mathrm{vec}\mathbf{V}_n)^\top)^\top$ is contained in $B_r(\boldsymbol{\theta})$. So, in order to apply Lemma 22.1, we place ourselves in the compact set $B_r(\boldsymbol{\theta})$. By applying Lemma 22.1 and Slutsky's theorem, we first prove that $\mathbf{A}_n(\boldsymbol{\mu}_n, \mathbf{V}_n)$ converges a.s. to $h(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = E(g(\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}))$ where:

$$g(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = K(\beta\|\mathbf{x} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top.$$

We have:

$$\begin{aligned}\|\mathbf{A}_n(\boldsymbol{\mu}_n, \mathbf{V}_n) - h(\boldsymbol{\mu}, \boldsymbol{\Sigma})\| \leq & \sup_{\boldsymbol{\nu}, \boldsymbol{\Gamma}} \|\mathbf{A}_n(\boldsymbol{\nu}, \boldsymbol{\Gamma}) - h(\boldsymbol{\nu}, \boldsymbol{\Gamma})\| \\ & + \|h(\boldsymbol{\mu}_n, \mathbf{V}_n) - h(\boldsymbol{\mu}, \boldsymbol{\Sigma})\|.\end{aligned}$$

Since K is bounded, $\|g(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\| \leq p_2(\|\mathbf{x}\|)$ where p_2 is a second degree polynomial. We assume that $\mathbf{X}$ has a second moment and so we can apply Lemma 22.1 and write:

$$\sup_{\boldsymbol{\nu}, \boldsymbol{\Gamma}} \|\mathbf{A}_n(\boldsymbol{\nu}, \boldsymbol{\Gamma}) - h(\boldsymbol{\nu}, \boldsymbol{\Gamma})\| \xrightarrow[n\to\infty]{\text{a.s.}} 0$$

Since g is bounded by an integrable function, h is continuous by Lebesgue dominated convergence theorem and $\mathbf{A}_n(\boldsymbol{\mu}_n, \mathbf{V}_n)$ converges a.s. to $h(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ by Slutsky's theorem.
In a similar way, we can prove that:

$$a_n(\boldsymbol{\mu}_n, \mathbf{V}_n) \xrightarrow[n\to\infty]{\text{a.s.}} E\left(K(\beta\|\mathbf{x} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})\right).$$

Finally, $\mathbf{S}_n(\beta)$ converges a.s. to $\mathbf{S}(\beta)$. □

In the gaussian case, with an exponential K function, we can give a more precise expression of Eq. (22.1) and derive the consistency of $\mathbf{U}_n(\beta)$.

Corollary 22.1 *If $F = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $K(x) = \exp(-x/2)$, then $\mathbf{S}_n(\beta)$ converges a.s. to $\frac{1}{1+\beta}\boldsymbol{\Sigma}$ and $\mathbf{U}_n(\beta)$ is a consistent estimator of $\boldsymbol{\Sigma}$.*

Note that under the regularity conditions given in Theorem 22.1 on the weight function K and for any elliptical distribution F, $\mathbf{U}_n(\beta)$ is a consistent estimator of $\boldsymbol{\Sigma}$ up to a multiplicative constant. In practice, we recommend the use of $K(x) = \exp(-x/2)$ which leads to a consistent estimator of $\boldsymbol{\Sigma}$ at the normal model.

22.2.3 *Asymptotic Distribution of* $\mathbf{U}_n(\beta)$

If the common distribution F has fourth moments and if the weight function K verifies some general regularity conditions, we can prove the asymptotic normality of $\mathbf{U}_n(\beta)$. In order to precise the asymptotic variances, we can consider an elliptical F distribution with fourth moments and an exponential weight function K. In this case, the detailed asymptotic variances for $\mathbf{U}_n(\beta)$ are given in Fekri and Ruiz-Gazen (2000) but will not be detailed in the present paper. In order to obtain explicit asymptotic variances, we focus on the gaussian case. Theorem 22.2 gives the asymptotic distribution of $\mathbf{U}_n(\beta)$ when F is gaussian and K is exponential.

Theorem 22.2 *If* $F = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $K(x) = \exp(-x/2)$ *and* $\sqrt{n}(\boldsymbol{\mu}_n - \boldsymbol{\mu})$ *converges in distribution, then* $\sqrt{n}\mathrm{vec}(\mathbf{U}_n(\beta) - \boldsymbol{\Sigma})$ *converges in distribution to a gaussian distribution with mean 0 and covariance matrix:*

$$\sigma_1(\mathbf{I}_{p^2} + \mathbf{C}_p)\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma} + \sigma_2 \mathrm{vec}\boldsymbol{\Sigma}(\mathrm{vec}\boldsymbol{\Sigma})^\top$$

with $\sigma_1 = (1+\beta)^{p+4}(1+2\beta)^{-\frac{p+4}{2}}$ *and* $\sigma_2 = \beta^2(1+\beta)^{-2}\sigma_1$.

Note that the proof is not trivial because the statistic $\mathbf{U}_n(\beta)$ is defined through the estimators $\boldsymbol{\mu}_n$ and $\mathbf{V}_n$.

Proof Let $\mathbf{m}$ denote a $p \times 1$ vector and $\boldsymbol{\Gamma}$ a $p \times p$ positive definite matrix.We define:

$$\mathbf{T}_n(\mathbf{m}, \boldsymbol{\Gamma}) = \frac{1}{n}\sum_{i=1}^{n} K(\beta \|\mathbf{X}_i - \mathbf{m}\|^2_{\boldsymbol{\Gamma}^{-1}})(\mathbf{X}_i - \mathbf{m}) \otimes (\mathbf{X}_i - \mathbf{m})$$

and

$$t_n(\mathbf{m}, \boldsymbol{\Gamma}) = \frac{1}{n}\sum_{i=1}^{n} K(\beta \|\mathbf{X}_i - \mathbf{m}\|^2_{\boldsymbol{\Gamma}^{-1}}).$$

We first prove that $\mathrm{vec}\mathbf{S}_n(\beta)$ and $\dfrac{\mathbf{T}_n(\boldsymbol{\mu}, \mathbf{V}_n)}{t_n(\boldsymbol{\mu}, \mathbf{V}_n)}$ have the same asymptotic distribution. We have:

$$\begin{aligned}
\mathrm{vec}\mathbf{S}_n(\beta) = {} & \frac{1}{t_n(\boldsymbol{\mu}_n, \mathbf{V}_n)}\sum_{i=1}^{n} K(\beta \|\mathbf{X}_i - \boldsymbol{\mu}_n\|^2_{\mathbf{V}_n^{-1}})(\mathbf{X}_i - \boldsymbol{\mu}) \otimes (\mathbf{X}_i - \boldsymbol{\mu}) \\
& -(\boldsymbol{\mu}_n - \boldsymbol{\mu}) \otimes \frac{1}{t_n(\boldsymbol{\mu}_n, \mathbf{V}_n)}\sum_{i=1}^{n} K(\beta \|\mathbf{X}_i - \boldsymbol{\mu}_n\|^2_{\mathbf{V}_n^{-1}})(\mathbf{X}_i - \boldsymbol{\mu}) \\
& -\frac{1}{t_n(\boldsymbol{\mu}_n, \mathbf{V}_n)}\sum_{i=1}^{n} K(\beta \|\mathbf{X}_i - \boldsymbol{\mu}_n\|^2_{\mathbf{V}_n^{-1}})(\mathbf{X}_i - \boldsymbol{\mu}) \otimes (\boldsymbol{\mu}_n - \boldsymbol{\mu}) \\
& +(\boldsymbol{\mu}_n - \boldsymbol{\mu}) \otimes (\boldsymbol{\mu}_n - \boldsymbol{\mu}).
\end{aligned}$$

We assume that $\sqrt{n}(\boldsymbol{\mu}_n - \boldsymbol{\mu})$ converges in distribution and we have:

$$\frac{1}{t_n(\boldsymbol{\mu}_n, \mathbf{V}_n)} \sum_{i=1}^{n} K(\beta \|\mathbf{X}_i - \boldsymbol{\mu}_n\|^2_{\mathbf{V}_n^{-1}})(\mathbf{X}_i - \boldsymbol{\mu})$$

which converges a.s. to:

$$\frac{E(K(\beta \|\mathbf{X} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})(\mathbf{X} - \boldsymbol{\mu}))}{E(K(\beta \|\mathbf{X} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}))}$$
$$= \frac{E(K(\beta \|\mathbf{X} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})\|\mathbf{X} - \boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}})}{E(K(\beta \|\mathbf{X} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}))} E\left(\frac{\mathbf{X} - \boldsymbol{\mu}}{\|\mathbf{X} - \boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}}\right) = 0.$$

So,

$$\text{vec}\mathbf{S}_n(\beta) = \frac{1}{t_n(\boldsymbol{\mu}_n, \mathbf{V}_n)} \sum_{i=1}^{n} K(\beta \|\mathbf{X}_i - \boldsymbol{\mu}_n\|^2_{\mathbf{V}_n^{-1}})(\mathbf{X}_i - \boldsymbol{\mu}) \otimes (\mathbf{X}_i - \boldsymbol{\mu}) + o\left(\frac{1}{\sqrt{n}}\right).$$

By using a Taylor-Lagrange expansion of:

$$\frac{1}{n} \sum_{i=1}^{n} K(\beta \|\mathbf{X}_i - \boldsymbol{\mu}_n\|^2_{\mathbf{V}_n^{-1}})(\mathbf{X}_i - \boldsymbol{\mu}) \otimes (\mathbf{X}_i - \boldsymbol{\mu})$$

at $\boldsymbol{\mu}_n = \boldsymbol{\mu}$, we obtain that:

$$\frac{1}{n} \sum_{i=1}^{n} K(\beta \|\mathbf{X}_i - \boldsymbol{\mu}_n\|^2_{\mathbf{V}_n^{-1}})(\mathbf{X}_i - \boldsymbol{\mu}) \otimes (\mathbf{X}_i - \boldsymbol{\mu}) = \mathbf{T}_n(\boldsymbol{\mu}, \mathbf{V}_n) + \xi_n(\tilde{\boldsymbol{\mu}}_n).(\boldsymbol{\mu}_n - \boldsymbol{\mu})$$

converges a.s. to 0 where

$$\xi_n(\tilde{\boldsymbol{\mu}}_n) = \frac{2\beta}{n} \sum_{i=1}^{n} K'(\beta \|\mathbf{X}_i - \tilde{\boldsymbol{\mu}}_n\|^2_{\mathbf{V}_n^{-1}})(\mathbf{X}_i - \boldsymbol{\mu}) \otimes (\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \tilde{\boldsymbol{\mu}}_n)^\top \mathbf{V}_n^{-1},$$

with $\tilde{\boldsymbol{\mu}}_n = \boldsymbol{\mu} + \alpha(\boldsymbol{\mu}_n - \boldsymbol{\mu})$ and $\alpha \in]0; 1[$. By Lemma 22.1,

$$\frac{1}{n} \sum_{i=1}^{n} K(\beta \|\mathbf{X}_i - \boldsymbol{\mu}_n\|^2_{\mathbf{V}_n^{-1}})(\mathbf{X}_i - \boldsymbol{\mu}) \otimes (\mathbf{X}_i - \boldsymbol{\mu}) = \mathbf{T}_n(\boldsymbol{\mu}, \mathbf{V}_n) + o\left(\frac{1}{\sqrt{n}}\right).$$

In the same way, we have:

$$t_n(\boldsymbol{\mu}_n, \mathbf{V}_n) = t_n(\boldsymbol{\mu}, \mathbf{V}_n) + o\left(\frac{1}{\sqrt{n}}\right),$$

so that:

$$\text{vec}\mathbf{S}_n(\beta) = \text{vec}\mathbf{S}(\beta) + \frac{1}{t}\left[\frac{t(\mathbf{T}_n(\boldsymbol{\mu}, \mathbf{V}_n) - \mathbf{T}) - (t_n(\boldsymbol{\mu}, \mathbf{V}_n) - t)\mathbf{T}}{t_n(\boldsymbol{\mu}_n, \mathbf{V}_n)}\right] + o\left(\frac{1}{\sqrt{n}}\right),$$

where

$$\mathbf{T} = \frac{1}{p}E(K(\beta\|\mathbf{X} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})\,\|\mathbf{X} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})\text{vec}\boldsymbol{\Sigma}$$

and

$$t = E(K(\beta\|\mathbf{X} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})).$$

Concerning $\mathbf{U}_n(\beta)$, we have:

$$\sqrt{n}(\mathbf{U}_n(\beta) - \boldsymbol{\Sigma}) = -\sqrt{n}(\mathbf{S}_n^{-1}(\beta) - \beta\mathbf{V}_n^{-1})^{-1}\left((\mathbf{S}_n^{-1}(\beta) - \mathbf{S}^{-1}(\beta)) - \beta(\mathbf{V}_n^{-1} - \boldsymbol{\Sigma}^{-1})\right)\boldsymbol{\Sigma}$$

which has the same asymptotic distribution as

$$\sqrt{n}\boldsymbol{\Sigma}\left(\beta(\mathbf{V}_n^{-1} - \boldsymbol{\Sigma}^{-1}) - (\mathbf{S}_n^{-1}(\beta) - \mathbf{S}^{-1}(\beta))\right)\boldsymbol{\Sigma}.$$

By using the vec notation, we have:

$$\begin{aligned}
&\sqrt{n}\text{vec}(\mathbf{U}_n(\beta) - \boldsymbol{\Sigma}) \\
&\quad = \sqrt{n}\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}\left(\mathbf{S}^{-1}(\beta) \otimes \mathbf{S}_n^{-1}(\beta)\text{vec}(\mathbf{S}_n(\beta) - \mathbf{S}(\beta))\right. \\
&\qquad \left. - \beta\boldsymbol{\Sigma}^{-1} \otimes \mathbf{V}_n^{-1}\text{vec}(\mathbf{V}_n - \boldsymbol{\Sigma})\right) + o(1) \\
&\quad = \sqrt{n}(1+\beta)^2\text{vec}(\mathbf{S}_n(\beta) - \mathbf{S}(\beta)) - \sqrt{n}\beta\text{vec}(\mathbf{V}_n - \boldsymbol{\Sigma}) + o(1) \\
&\quad = \sqrt{n}\left(\frac{1+\beta}{t}\right)^2 (t(\mathbf{T}_n(\boldsymbol{\mu}, \mathbf{V}_n) - \mathbf{T}) - (t_n(\boldsymbol{\mu}, \mathbf{V}_n) - t)\mathbf{T}) \\
&\qquad -\sqrt{n}\beta\text{vec}(\mathbf{V}_n - \boldsymbol{\Sigma}) + o(1) \\
&\quad = \sqrt{n}\left(\frac{1+\beta}{t}\right)^2 (t(\mathbf{T}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) - \mathbf{T}) + t\Delta_n\text{vec}(\mathbf{V}_n - \boldsymbol{\Sigma}) \\
&\qquad - ((t_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) - t) + \delta_n\text{vec}(\mathbf{V}_n - \boldsymbol{\Sigma}))\mathbf{T}) - \sqrt{n}\beta\text{vec}(\mathbf{V}_n - \boldsymbol{\Sigma}) + o(1) \\
&\quad = \sqrt{n}\left(\frac{(1+\beta)^2}{t}\mathbf{I}_{p^2}, -\left(\frac{1+\beta}{t}\right)^2\mathbf{T}, \mathbf{B}_n\right)\begin{pmatrix}\mathbf{T}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) - \mathbf{T} \\ t_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) - t \\ \text{vec}(\mathbf{V}_n - \boldsymbol{\Sigma})\end{pmatrix} + o(1)
\end{aligned}$$

where

$$\mathbf{B}_n = \frac{(1+\beta)^2}{t}\Delta_n - \left(\frac{1+\beta}{t}\right)^2 \mathbf{T}\,\delta_n - \beta \mathbf{I}_{p^2},$$

$$\begin{aligned}\Delta_n &= \left.\frac{\partial \mathbf{T}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{(\partial \mathrm{vec}\,\boldsymbol{\Sigma})^\top}\right|_{\tilde{\mathbf{V}}_n} \\ &= -\frac{\beta}{n}\sum_{i=1}^{n} K'(\beta \|\mathbf{X}_i - \boldsymbol{\mu}\|^2_{\widetilde{\mathbf{V}}_n^{-1}})\,\mathrm{vec}(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})^\top \\ &\qquad \left(\mathrm{vec}(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})^\top\right)^\top \widetilde{\mathbf{V}}_n^{-1} \otimes \widetilde{\mathbf{V}}_n^{-1},\end{aligned}$$

and

$$\delta_n = \left.\frac{\partial t_n}{(\partial \mathrm{vec}\,\boldsymbol{\Sigma})^\top}\right|_{\tilde{\mathbf{V}}_n} = -\frac{\beta}{n}\sum_{i=1}^{n} K'(\|\mathbf{X} - \boldsymbol{\mu}\|^2_{\widetilde{\mathbf{V}}_n^{-1}}) \left(\mathrm{vec}(\widetilde{\mathbf{V}}_n^{-1}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top \widetilde{\mathbf{V}}_n^{-1}\right)^\top.$$

By the central limit theorem, the associated projector

$$\sqrt{n}(\mathbf{T}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) - \mathbf{T}, t_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) - t, \mathrm{vec}(\mathbf{V}_n - \boldsymbol{\Sigma}))^\top$$

converges to a gaussian distribution with zero mean and covariance matrix:

$$\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\Theta}_{11} & \boldsymbol{\Theta}_{12} & \boldsymbol{\Theta}_{13} \\ \boldsymbol{\Theta}_{12} & \boldsymbol{\Theta}_{22} & \boldsymbol{\Theta}_{23} \\ \boldsymbol{\Theta}_{13} & \boldsymbol{\Theta}_{23} & \boldsymbol{\Theta}_{33} \end{pmatrix}$$

where

$$\boldsymbol{\Theta}_{11} = \frac{k_3(2\beta)}{p(p+2)}(\mathbf{I}_{p^2} + \mathbf{C}_p)\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma} + \left(\frac{k_3(2\beta)}{p(p+2)} - \frac{k_2^2(\beta)}{p^2}\right)\mathrm{vec}\,\boldsymbol{\Sigma}\,(\mathrm{vec}\,\boldsymbol{\Sigma})^\top,$$

$$\boldsymbol{\Theta}_{12} = \frac{k_2(2\beta) - k_1(\beta)k_2(\beta)}{p}(\mathrm{vec}\,\boldsymbol{\Sigma})^\top,$$

$$\boldsymbol{\Theta}_{22} = k_1(2\beta) - k_1^2(\beta),$$

$$\boldsymbol{\Theta}_{13} = \frac{k_3(\beta)}{p(p+2)}(\mathbf{I}_{p^2} + \mathbf{C}_p)\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma} + \left(\frac{k_3(\beta)}{p(p+2)} - \frac{k_2(\beta)}{p}\right)\mathrm{vec}\,\boldsymbol{\Sigma}\,(\mathrm{vec}\,\boldsymbol{\Sigma})^\top,$$

$$\boldsymbol{\Theta}_{23} = \left(\frac{k_2(\beta)}{p} - k_1(\beta)\right)\mathrm{vec}\,\boldsymbol{\Sigma},$$

$$\boldsymbol{\Theta}_{33} = (\mathbf{I}_{p^2} + \mathbf{C}_p)\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}$$

with

$$k_1(\beta) = E\left[K(\beta\|\mathbf{X}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})\right],$$
$$k_2(\beta) = E\left[K(\beta\|\mathbf{X}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})\|\mathbf{X}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}\right],$$
$$k_3(\beta) = E[K(\beta\|\mathbf{X}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})\|\mathbf{X}-\boldsymbol{\mu}\|^4_{\boldsymbol{\Sigma}^{-1}}]$$

and $K(x) = \exp(-x/2)$.

$\mathbf{B}_n$ converges a.s. to:

$$\begin{aligned}\mathbf{B} &= -\frac{\beta(1+\beta)^2}{p(p+2)t}E\left[K'(\beta\,\|\mathbf{X}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})\|\mathbf{X}-\boldsymbol{\mu}\|^4_{\boldsymbol{\Sigma}^{-1}}\right]\left(\mathbf{I}_{p^2}+\mathbf{C}_p+\mathrm{vec}\,\boldsymbol{\Sigma}(\mathrm{vec}\,\boldsymbol{\Sigma}^{-1})^\top\right)\\ &\quad+\frac{\beta(1+\beta)^2}{p\,t^2}E\left[K'(\beta\,\|\mathbf{X}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}))\|\mathbf{X}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}\right]\mathbf{T}\,\left(\mathrm{vec}\,\boldsymbol{\Sigma}^{-1}\right)^\top-\beta\mathbf{I}_{p^2}\\ &= \frac{\beta(1+\beta)^2}{2p(p+2)}\frac{k_3(\beta)}{k_1(\beta)}\left(I_{p^2}+C_p+\mathrm{vec}\,\boldsymbol{\Sigma}(\mathrm{vec}\,\boldsymbol{\Sigma}^{-1})^\top\right)\\ &\quad-\frac{\beta(1+\beta)^2}{2p^2}\frac{k_2^2(\beta)}{k_1^2(\beta)}\mathrm{vec}\,\boldsymbol{\Sigma}\,\left(\mathrm{vec}\,\boldsymbol{\Sigma}^{-1}\right)^\top-\beta\mathbf{I}_{p^2}.\end{aligned}$$

So, we obtain that $\sqrt{n}\mathrm{vec}(\mathbf{U}_n(\beta)-\boldsymbol{\Sigma})$ converges to a gaussian distribution with zero mean and covariance matrix:

$$\begin{aligned}&\left(\frac{(1+\beta)^2}{t}\mathbf{I}_{p^2},-\left(\frac{1+\beta}{t}\right)^2\mathbf{T},\,\mathbf{B}\right)\boldsymbol{\Theta}\left(\frac{(1+\beta)^2}{t}\mathbf{I}_{p^2},-\left(\frac{1+\beta}{t}\right)^2\mathbf{T},\,\mathbf{B}\right)^\top\\ &\qquad=(1+\beta)^{p+4}(1+2\beta)^{-\frac{p+4}{2}}(\mathbf{I}_{p^2}+\mathbf{C}_p)\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}\\ &\qquad\quad+(1+\beta)^{p+4}(1+2\beta)^{-\frac{p+4}{2}}\beta^2\mathrm{vec}\,\boldsymbol{\Sigma}\,(\mathrm{vec}\,\boldsymbol{\Sigma})^\top\end{aligned}$$

□

22.2.4 *Asymptotic Distribution of the Eigenelements of* $\mathbf{U}_n(\boldsymbol{\beta})$

Now, we are interested in the asymptotic distributions of the eigenvalues and eigenvectors of $\mathbf{U}_n(\beta)$. We use results by Dauxois et al. (1982), Dossou-Gbete and Pousse (1991), and Fekri and Fine (1995) on the convergence of eigenelements of a random matrix.

Let us consider λ_i an eigenvalue of $\boldsymbol{\Sigma}$ with multiplicity k_i and $\mathbf{P}_i$ the associated projector. We can find k_i series of eigenvalues $\hat{\lambda}_{ij}$ of $\mathbf{U}_n(\beta)$ which converge a.s. to

λ_i. We define:

$$\hat{\lambda}_i = \frac{1}{k_i} \sum_{j=1}^{k_i} \hat{\lambda}_{ij}$$

and $\hat{\mathbf{P}}_i$ the sum of the projectors associated with the k_i eigenvalues.

Theorem 22.3 *If* $F = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $K(x) = \exp(-x/2)$ *and* $\sqrt{n}(\boldsymbol{\mu}_n - \boldsymbol{\mu})$ *converges in distribution, then:*

(i) $\sqrt{n}(\hat{\lambda}_i - \lambda_i)$ *converges to a gaussian distribution with mean 0 and variance:*

$$\lambda_i^2 \left(\frac{2\sigma_1}{k_i} + \sigma_2 \right),$$

(ii) $\sqrt{n}\mathrm{vec}(\hat{\mathbf{P}}_i - \mathbf{P}_i)$ *converges to a gaussian distribution with mean 0 and covariance matrix:*

$$\sigma_1 (\mathbf{I}_{p^2} + C_p) \sum_{j \neq i} \frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} (\mathbf{P}_i \otimes \mathbf{P}_j + \mathbf{P}_j \otimes \mathbf{P}_i).$$

Proof Dauxois et al. (1982), Dossou-Gbete and Pousse (1991), Fekri and Fine (1995) have studied the asymptotic distribution of the eigenelements of a random matrix. So, it is known that if

$$\sqrt{n}(\mathbf{U}_n(\beta) - \boldsymbol{\Sigma})$$

converges in distribution to a gaussian variable $\mathbf{Y}$ with mean zero then

$$\sqrt{n}(\hat{\lambda}_i - \lambda_i)$$

converges in distribution to

$$\frac{1}{k_i} \mathrm{tr}(\mathbf{P}_i \mathbf{Y})$$

a gaussian variable with mean zero and covariance matrix:

$$\frac{1}{k_i^2} E[(\mathrm{vec}\mathbf{P}_i)^\top \mathrm{vec}\mathbf{Y} (\mathrm{vec}\mathbf{Y})^\top \mathrm{vec}\mathbf{P}_i]$$

$$= \frac{1}{k_i^2} (\mathrm{vec}\mathbf{P}_i)^\top \left(\sigma_1 (\mathbf{I}_{p^2} + C_p) \boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma} + \sigma_2 \mathrm{vec}\boldsymbol{\Sigma} (\mathrm{vec}\boldsymbol{\Sigma})^\top \right) \mathrm{vec}\mathbf{P}_i$$

$$= \lambda_i^2 \left(2\frac{\sigma_1}{k_j} + \sigma_2 \right).$$

We also have that $\sqrt{n}\text{vec}(\hat{\mathbf{P}}_i - \mathbf{P}_i)$ converges in distribution to

$$\text{vec}(\mathbf{S}_i\mathbf{Y}\mathbf{P}_i + \mathbf{P}_i\mathbf{Y}\mathbf{S}_i),$$

a gaussian variable with mean zero and covariance matrix:

$$\begin{aligned}
&E[(\mathbf{P}_i \otimes \mathbf{S}_i + \mathbf{S}_i \otimes \mathbf{P}_i)\text{vec}\mathbf{Y}(\text{vec}\mathbf{Y})^\top(\mathbf{P}_i \otimes \mathbf{S}_i + \mathbf{S}_i \otimes \mathbf{P}_i) \\
&\quad = \sigma_1(\mathbf{I}_{p^2} + \mathbf{C}_p)\,(\mathbf{P}_i\boldsymbol{\Sigma}\mathbf{S}_i \otimes \mathbf{S}_i\boldsymbol{\Sigma}\mathbf{P}_i + \mathbf{P}_i\boldsymbol{\Sigma}\mathbf{P}_i \otimes \mathbf{S}_i\boldsymbol{\Sigma}\mathbf{S}_i + \mathbf{S}_i\boldsymbol{\Sigma}\mathbf{S}_i \otimes \mathbf{P}_i\boldsymbol{\Sigma}\mathbf{P}_i \\
&\qquad + \mathbf{S}_i\boldsymbol{\Sigma}\mathbf{P}_i \otimes \mathbf{P}_i\boldsymbol{\Sigma}\mathbf{S}_i) \\
&\quad = \sigma_1\lambda_i(\mathbf{I}_{p^2} + \mathbf{C}_p)\,(\mathbf{P}_i \otimes \mathbf{S}_i\boldsymbol{\Sigma}\mathbf{S}_i + \mathbf{S}_i\boldsymbol{\Sigma}\mathbf{S}_i \otimes \mathbf{P}_i) \\
&\quad = \sigma_1(\mathbf{I}_{p^2} + \mathbf{C}_p)\left(\sum_{j\neq i}\frac{\lambda_i\lambda_j}{(\lambda_i - \lambda_j)^2}(\mathbf{P}_i \otimes \mathbf{P}_j + \mathbf{P}_j \otimes \mathbf{P}_i)\right).
\end{aligned}$$

□

If λ_i, λ_j are two *simple* eigenvalues of $\boldsymbol{\Sigma}$ and $\mathbf{u}_i, \mathbf{u}_j$ are the associated eigenvectors respectively, we can prove the following corollary:

Corollary 22.2 *If $F = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $K(x) = \exp(-x/2)$ and $\sqrt{n}(\boldsymbol{\mu}_n - \boldsymbol{\mu})$ converges in distribution, then:*

(i) $\sqrt{n}\begin{pmatrix}\widehat{\lambda_i} - \lambda_i \\ \widehat{\lambda_j} - \lambda_j\end{pmatrix}$ *converges to a gaussian distribution with mean 0 and covariance matrix:*

$$\begin{pmatrix}\lambda_i^2(2\sigma_1 + \sigma_2) & \lambda_i\lambda_j\sigma_2 \\ \lambda_i\lambda_j\sigma_2 & \lambda_j^2(2\sigma_1 + \sigma_2)\end{pmatrix},$$

(ii) $\sqrt{n}\begin{pmatrix}\widehat{\mathbf{u}_i} - \mathbf{u}_i \\ \widehat{\mathbf{u}_j} - \mathbf{u}_j\end{pmatrix}$ *converges to a gaussian distribution with mean 0 and covariance matrix:*

$$\sigma_1\begin{pmatrix}\displaystyle\sum_{k\neq i}\frac{\lambda_i\lambda_k}{(\lambda_i - \lambda_k)^2}\mathbf{u}_k\mathbf{u}_k^\top & -\dfrac{\lambda_i\lambda_j}{(\lambda_i - \lambda_j)^2}\mathbf{u}_j\mathbf{u}_i^\top \\ -\dfrac{\lambda_i\lambda_j}{(\lambda_i - \lambda_j)^2}\mathbf{u}_i\mathbf{u}_j^\top & \displaystyle\sum_{k\neq j}\frac{\lambda_j\lambda_k}{(\lambda_j - \lambda_k)^2}\mathbf{u}_k\mathbf{u}_k^\top\end{pmatrix}.$$

The asymptotic variances of the eigenvalues may be used to construct large sample confidence intervals and tests. Croux and Haesbroeck (2000) give the asymptotic efficiencies at normal distributions for a simple eigenvalue λ_i estimator and for its associated eigenvector estimator v_i based on the M-, the S-, and the one-step reweighted Minimum Covariance Determinant (RMCD) estimators which are

Table 22.1 Asymptotic efficiencies of the eigenelements of $\mathbf{U}_n(\beta)$, M, S, and RMCD at the gaussian model

		Dimensions				
		$p = 2$	$p = 3$	$p = 5$	$p = 10$	$p = 30$
Eff(λ_i)	$\mathbf{U}_n(\beta)$	0.665	0.774	0.869	0.905	0.970
	M	0.881	0.895	0.947	0.974	0.991
	S	0.895	0.941	0.968	0.990	0.997
	RMCD	0.599	0.680	0.753	0.836	0.901
Eff(v_i)	$\mathbf{U}_n(\beta)$	0.702	0.798	0.881	0.911	0.971
	M	0.920	0.947	0.969	0.986	0.996
	S	0.850	0.924	0.967	0.988	0.997
	RMCD	0.635	0.742	0.820	0.873	0.933

classical robust covariance estimators. In the present paper, we do not recall the definition and the properties of these estimators but the reader can refer to Croux and Haesbroeck (2000) for a survey. In order to compare the efficiencies at normal distributions for eigenelements estimators based on $\mathbf{U}_n(\beta)$ with estimators based on the M-, the S-, and the RMCD estimators, we consider (Table 22.1) the same dimension values as in the second table of Croux and Haesbroeck (2000). We recall all the values from Croux and Haesbroeck (2000) and add the values for the $\mathbf{U}_n(\beta)$ estimator. We denote by Eff these efficiencies. We have $\text{Eff}(\lambda_i) = 2/(2\sigma_1 + \sigma_2)$ and $\text{Eff}(v_i) = 1/\sigma_1$ if we choose the Hilbert-Schmidt norm for the asymptotic covariance matrix of v_i. Following the recommendation given in Sect. 22.2.1, we take $\beta = 1/p$ for $p = 2, 3, 5$ and $\beta = (26\varepsilon)/p$ with $\varepsilon = 5\,\%$ for $p = 10, 30$. Definitions of the M-, S-, and RMCD estimators also involve choices of functions and constants which are detailed in Croux and Haesbroeck (2000). The efficiencies for $\mathbf{U}_n(\beta)$ are relatively high but are not as good as the ones obtained for the M- and the S-estimators by Croux and Haesbroeck. Nevertheless, they are better than the efficiencies based on the RMCD estimator.

22.3 Influence Function of $\mathbf{U}_n(\beta)$ and of Its Eigenelements

The influence function of an estimator measures the impact of an infinitesimal data contamination on the estimator (see Hampel et al. 1986 for more details). The influence function of a robust estimator is expected to be bounded and such estimators are said B-robust.

In Ruiz-Gazen (1996), we calculate the influence function of $\mathbf{U}_n(\beta)$ and obtain that, if $F = N(0, \boldsymbol{\Sigma})$, $K(x) = \exp(-x/2)$, and $\boldsymbol{\mu}_n$ is an affine equivariant estimator, then the influence function of the functional $\mathbf{U}$ associated with $\mathbf{U}_n$ is:

$$IF(\mathbf{x}, \mathbf{U}, F) = \alpha_1(\|\mathbf{x} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top - \alpha_2(\|\mathbf{x} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})\boldsymbol{\Sigma}$$

with

$$\alpha_1(a) = (1+\beta)^{\frac{p+4}{2}} K(\beta a)$$

and

$$\alpha_2 = (1+\beta)\alpha_1.$$

Since the functions $x \mapsto x\alpha_1(x)$ and α_2 are bounded for $\beta > 0$, the influence function of $\mathbf{U}_n(\beta)$ is bounded and $\mathbf{U}_n(\beta)$ is B-robust.

Now, we give the influence function of the eigenelements of $\mathbf{U}_n(\beta)$. We do not need any assumption concerning the convergence of $\boldsymbol{\mu}_n$ but just assume that $\boldsymbol{\mu}_n$ is an affine equivariant estimator.

Theorem 22.4 *If $\boldsymbol{\mu}_n$ is an affine equivariant estimator, then*

$$I\!F(\mathbf{x}, \lambda_i; F) = \frac{\alpha_1(\|\mathbf{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})}{k_i}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{P}_i(\mathbf{x}-\boldsymbol{\mu}) - \alpha_2(\|\mathbf{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})\lambda_i$$

$$I\!F(\mathbf{x}, \mathbf{P}_i; F) = \alpha_1(\|\mathbf{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}) \left(\mathbf{S}_i(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{P}_i + \mathbf{P}_i(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{S}_i\right)$$

with

$$\mathbf{S}_i = \sum_{j \neq i} \frac{1}{\lambda_i - \lambda_j} \mathbf{P}_j.$$

Proof We have:

$$\mathbf{U}((1-\varepsilon)F + \varepsilon \Delta_{\mathbf{x}}) = \mathbf{U}(F) + \varepsilon I\!F(\mathbf{x}, \mathbf{U}; F) + O(\varepsilon^2),$$

where $\Delta_{\mathbf{x}}$ denotes the Dirac distribution at $\mathbf{x}$. By the perturbation theory of linear operators (Kato 1966) we can write:

$$\lambda_j((1-\varepsilon)F + \varepsilon \Delta_{\mathbf{x}}) = \lambda_i(F) + \varepsilon \gamma_j + O(\varepsilon^2)$$

for all $j \in J_i$, where $J_i = \{k, \lambda_k = \lambda_i\}$ and $(\gamma_j)_{j \in J_i}$ is the full decreasing series of the eigenvectors of $\mathbf{P}_i I\!F(\mathbf{x}, \mathbf{U}; F)\mathbf{P}_i$.
If $\lambda_i((1-\varepsilon)F + \varepsilon \Delta_{\mathbf{x}})$ is the mean of the k_i eigenvalues of $\mathbf{U}((1-\varepsilon)F + \varepsilon \Delta_{\mathbf{x}})$, then:

$$\lambda_i((1-\varepsilon)F + \varepsilon \Delta_{\mathbf{x}}) = \lambda_i(F) + \frac{\varepsilon}{k_i} \mathrm{tr}(\mathbf{P}_i I\!F(\mathbf{x}, \mathbf{U}; F)) + O(\varepsilon^2)$$

It follows that:

$$I\!F(\mathbf{x}, \lambda_i; F) = \frac{1}{k_i} \mathrm{tr}(\mathbf{P}_i I\!F(\mathbf{x}, \mathbf{U}; F))$$

and replacing $IF(\mathbf{x}, \mathbf{U}; F)$ by its expression yields to:

$$IF(\mathbf{x}, \lambda_i; F) = \frac{\alpha_1(\|\mathbf{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})}{k_i}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{P}_i(\mathbf{x}-\boldsymbol{\mu}) - \alpha_2(\|\mathbf{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})\lambda_i.$$

For the eigenprojectors,

$$\mathbf{P}_i((1-\varepsilon)F + \varepsilon\Delta_{\mathbf{x}}) = \mathbf{P}_i(F) + \varepsilon(\mathbf{P}_i IF(\mathbf{x}, \mathbf{U}; F)\mathbf{S}_i + \mathbf{S}_i IF(\mathbf{x}, \mathbf{U}; F)\mathbf{P}_i) + O(\varepsilon^2)$$

and because $\mathbf{P}_i\boldsymbol{\Sigma}\mathbf{S}_i = 0$,

$$IF(\mathbf{x}, \mathbf{P}_i; F) = \alpha_1(\|\mathbf{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})\left(\mathbf{S}_i(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{P}_i + \mathbf{P}_i(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{S}_i\right).$$

□

Corollary 22.3 *If $\boldsymbol{\mu}_n$ is an affine equivariant estimator, and λ_i is a* simple *eigenvalue of $\boldsymbol{\Sigma}$ associated with the normalized eigenvector u_i, then:*

$$IF(\mathbf{x}, \lambda_i; F) = \alpha_1(\|\mathbf{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})((\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{u}_i)^2 - \alpha_2(\|\mathbf{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})\lambda_i$$
$$IF(\mathbf{x}, u_i; F) = \alpha_1(\|\mathbf{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{u}_i\mathbf{S}_i(\mathbf{x}-\boldsymbol{\mu}).$$

Note that similar results are obtained by Critchley (1985) for simple eigenvalues and by Tanaka (1988) for multiple eigenvalues. In this paper, proofs are simplified by using the perturbation theory of linear operators (Kato 1966). Since $x \mapsto x\alpha_1(x)$ and α_2 are bounded functions, the influence function of the eigenelements of $\mathbf{U}_n(\beta)$ are bounded. So, $\mathbf{U}_n(\beta)$ has good infinitesimal robustness properties. These results are confirmed in the following sections when considering simulations and a real data set.

22.4 Simulations and Comparisons

In this section, we use a similar simulation scheme as in Devlin et al. (1981) in order to compare the performance of $\mathbf{U}_n(\beta)$ with other robust and non-robust estimators. The parameter to estimate is not a covariance matrix but a *correlation* matrix $\boldsymbol{\Lambda}$. If $\boldsymbol{\Sigma}$ denotes the population covariance matrix, we have $\boldsymbol{\Lambda} = (\text{diag}(\boldsymbol{\Sigma}))^{-1/2}\boldsymbol{\Sigma}(\text{diag}(\boldsymbol{\Sigma}))^{-1/2}$. If $\hat{\boldsymbol{\Sigma}}$ is an estimator of $\boldsymbol{\Sigma}$, then

$$\hat{\boldsymbol{\Lambda}} = (\text{diag}(\hat{\boldsymbol{\Sigma}}))^{-1/2}\hat{\boldsymbol{\Sigma}}(\text{diag}(\hat{\boldsymbol{\Sigma}}))^{-1/2}$$

is an estimator of $\boldsymbol{\Lambda}$. Thus a correlation estimator is naturally associated with a covariance estimator by the previous transformation. In the following, we use the same notation for the covariance estimator and its associated correlation estimator. We compare the $\mathbf{U}_n(\beta)$ estimator with the classical non-robust correlation estimator and also with the Huber M-estimator and the biweight S-estimator which are well-known robust estimators. A detailed comparison between M-, S-, and also RMCD estimators is given in Croux and Haesbroeck (2000) according to the same simulation scheme. In the present paper, we use the procedures they propose to calculate the M- and S- estimators. More precisely, we use Huber's proposal for the weight functions of the M-estimator and the biweight function for the S-estimator with a constant chosen such that the breakdown point of the S-estimator is 25 %. We denote by R the usual non-robust correlation estimator which is defined by:

$$(\text{diag}(\mathbf{V}_n))^{-1/2}\mathbf{V}_n(\text{diag}(\mathbf{V}_n))^{-1/2}.$$

We denote by M the Huber M-estimator, by S the biweight S-estimator, and by U the estimator we propose. On this particular simulation scheme, we found that calculating U is 3 times faster than calculating the Huber M-estimator and more than 20 times faster than calculating the biweight S-estimator.

We generate $m = 1000$ replications of samples of size $n = 50$ from several distributions of dimension $p = 6$. For each distribution, we use the correlation matrix $\boldsymbol{\Lambda}$ defined in Devlin et al. (1981) by:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_1 & 0 \\ 0 & \boldsymbol{\Lambda}_2 \end{bmatrix} \text{ with } \boldsymbol{\Lambda}_1 = \begin{bmatrix} 1 & & \\ .95 & 1 & \\ .30 & .10 & 1 \end{bmatrix} \text{ and } \boldsymbol{\Lambda}_2 = \begin{bmatrix} 1 & & \\ -.499 & 1 & \\ -.499 & -.499 & 1 \end{bmatrix}.$$

In the following, we denote by ρ the elements of $\boldsymbol{\Lambda}$. The eigenvalues of $\boldsymbol{\Lambda}$ are $\lambda_1 = 2.029$, $\lambda_2 = \lambda_3 = 1.499$, $\lambda_4 = .943$, $\lambda_5 = .028$, $\lambda_6 = .002$ and the eigenvector corresponding to λ_6 is $v_6^\top = 1/\sqrt{3}(0, 0, 0, 1, 1, 1)$. The matrix $\boldsymbol{\Lambda}$ has some desirable features that have been described in Devlin et al. (1981). We are interested in the estimation of the correlation coefficients, eigenvalues and eigenvectors of $\boldsymbol{\Lambda}$. We consider samples from five distributions:

1. the normal distribution: $\mathcal{N}(0, \boldsymbol{\Lambda})$ denoted by NOR$[0, \boldsymbol{\Lambda}]$,
2. the mixture 0.9 NOR$[0, \boldsymbol{\Lambda}]$ + 0.1 NOR$[0, 9\boldsymbol{\Lambda}]$ which is a symmetric contaminated normal distribution denoted by SCN$[0, \boldsymbol{\Lambda}]$,
3. the Cauchy distribution denoted by CAU$[0, \boldsymbol{\Lambda}]$,
4. the mixture 0.9 NOR$[0, \boldsymbol{\Lambda}]$ + 0.1 NOR$[\boldsymbol{\mu}_1, \boldsymbol{\Lambda}]$ with $\boldsymbol{\mu}_1 = .537v_6$ which is an asymmetric contaminated normal distribution denoted by ACN1$[\boldsymbol{\mu}_1, \boldsymbol{\Lambda}]$.
5. the mixture 0.9 NOR$[0, \boldsymbol{\Lambda}]$ + 0.1 NOR$[\boldsymbol{\mu}_2, \boldsymbol{\Lambda}]$ with $\boldsymbol{\mu}_2 = 4.472v_6$ which is an asymmetric contaminated normal distribution denoted by ACN2$[\boldsymbol{\mu}_2, \boldsymbol{\Lambda}]$.

In the ACN1 (respectively ACN2) distribution, the 10 % contaminated data have been shifted about 12 (respectively 100) standard deviation units out along the direction of v_6. Devlin et al. (1981) only consider the ACN1 asymmetric distribution. In the present paper, we also consider ACN2 to model more extreme outliers. We have to choose a location estimator $\boldsymbol{\mu}_n$ and a value of β in order to calculate $\mathbf{U}_n(\beta)$. The location estimator we use is the spatial median calculated with the algorithm of Bedall and Zimmerman (1979). This estimator is orthogonal equivariant, has a high breakdown point, and can be computed very fast. A possible alternative to this robust location estimator is the so-called Oja median (Oja 1983) which is affine equivariant and is implemented in the R package OjaNP (Fischer et al. 2013). Concerning β, we propose to use $\beta_0 = (26\varepsilon)/p$ with $\varepsilon = 10\,\%$ for the five distributions. Thus, $\beta_0 = .4$. We calculate biases and mean squared errors (MSE) for the correlations and the eigenvalues estimates and cosines for the eigenvectors estimates.

22.4.1 Correlations and Eigenvalues Estimators

The Monte Carlo biases and the MSE of the Fisher's z transform of the estimates of the elements of $\boldsymbol{\Lambda}$ are given in Table 22.2 while the biases and MSE of the Logs of the estimated eigenvalues are given in Table 22.3. We recall that for an estimator $\hat{\theta}$ of a parameter θ, if we denote by $\hat{\theta}^{(k)}$ the estimate value for replication $k = 1, \ldots, m$, the Monte Carlo bias of $\hat{\theta}$ is defined by:

$$\text{Bias} = \frac{1}{m}\sum_{k=1}^{m}\left(\hat{\theta}^{(k)} - \theta\right)$$

and the Monte Carlo MSE is:

$$\text{MSE} = \frac{1}{m}\sum_{k=1}^{m}(\hat{\theta}^{(k)} - \theta)^2.$$

In Table 22.2, the MSEs are calculated for the Fisher's z transform of the elements of the matrix estimators which is defined by:

$$f(z) = \frac{1}{2}\ln\left(\frac{1+z}{1-z}\right).$$

Note that the bias and the MSE are not given for all the zeros of the $\boldsymbol{\Lambda}$ matrix. Only the largest bias and MSE are reported in the "0-max" rows.

For the normal distribution, the four estimates give similar results with a bias slightly superior for the $\mathbf{U}_n(\beta)$ estimator. For the contaminated models, robust

Table 22.2 1000× Biases and 1000× MSE of the z-transform for the estimators R, M, S, U of the ρ elements of $\boldsymbol{\Lambda}$

		1000× Bias				1000× MSE of z-transform			
	$1000 \times \rho$	R	M	S	U	R	M	S	U
NOR	950	0	0	0	−1	21	21	21	29
	300	−4	−4	−4	−10	20	20	20	30
	100	−1	−1	−2	−7	19	20	20	30
	−499	4	4	4	4	22	22	22	32
	−499	3	3	3	4	22	22	23	32
	−499	5	5	5	8	20	20	21	31
	0-max	8	8	7	7	22	22	23	34
SCN	950	−2	0	0	−1	53	23	27	27
	300	−7	−5	−7	−8	52	23	27	29
	100	−3	−2	−3	−4	53	23	27	27
	−499	13	6	6	3	56	26	29	30
	−499	6	2	4	5	55	25	29	31
	−499	9	5	5	8	53	23	27	29
	0-max	7	9	12	8	57	25	29	31
CAU	950	−47	0	−2	−5	1425	37	63	124
	300	−35	−8	−10	−10	1535	34	57	128
	100	3	−3	−5	−2	1469	33	57	114
	−499	103	6	8	13	1538	36	61	106
	−499	41	4	6	2	1531	39	61	123
	−499	66	9	17	16	1478	36	63	107
	0-max	−21	8	−11	−13	1538	41	61	126
ACN1	950	0	0	0	−1	21	21	21	30
	300	−4	−4	−4	−8	20	21	21	32
	100	−1	−2	−2	−5	19	20	21	31
	−499	18	14	12	9	21	22	22	32
	−499	15	13	12	11	22	23	23	33
	−499	18	14	12	11	20	21	21	31
	0-max	7	−8	8	6	22	23	24	36
ACN2	950	0	0	0	−1	21	22	23	31
	300	−4	−4	−4	−9	20	21	23	32
	100	−1	−2	−2	−5	19	20	22	31
	−499	576	405	56	176	408	226	53	67
	−499	571	401	55	179	402	223	54	69
	−499	575	406	57	181	406	228	53	70
	0-max	−7	−8	8	−6	22	22	25	33

Table 22.3 1000× Biases and 100× MSE of the Logs of the eigenvalues estimators based on the R, M, S, and U correlation matrix estimators

		1000× Bias				100× MSE of Logs			
	λ	R	M	S	U	R	M	S	U
NOR	2.029	189	190	195	196	1	1	1	1
	1.499	111	111	111	111	1	1	1	1
	1.499	−190	−190	−194	−195	3	3	1	3
	0.943	−108	−109	−111	−111	3	3	3	3
	0.028	−2	−2	−2	−2	9	9	9	9
	0.002	0	0	0	0	9	9	9	9
SCN	2.029	403	213	237	251	4	1	2	2
	1.499	137	115	120	119	2	1	1	1
	1.499	−313	−203	−217	−227	8	3	4	4
	0.943	−223	−123	−137	−141	13	4	5	6
	0.028	−3	−2	−2	−2	24	11	12	12
	0.002	0	0	0	0	23	10	12	12
CAU	2.029	1830	302	427	391	44	3	5	5
	1.499	−149	134	162	125	148	1	2	10
	1.499	−939	−260	−326	−299	530	5	29	29
	0.943	−724	−172	−259	−213	935	8	19	41
	0.028	−17	−3	−4	−3	873	17	31	50
	0.002	−2	0	0	0	1150	17	31	54
ACN1	2.029	186	195	199	273	1	1	1	2
	1.499	101	105	105	115	1	1	1	1
	1.499	−201	−204	−203	−242	3	3	3	4
	0.943	−109	−113	−114	−153	3	4	4	6
	0.028	3	1	0	−2	5	5	6	12
	0.002	19	16	13	8	555	481	394	269
ACN2	2.029	134	134	203	224	1	1	1	2
	1.499	−229	−197	76	−20	4	3	1	1
	1.499	−457	−454	−235	−362	14	14	5	9
	0.943	−94	−108	−116	−170	2	3	4	7
	0.028	623	601	70	304	975	953	119	597
	0.002	24	24	2	24	648	651	80	644

estimators clearly outperform the usual non-robust estimator. As it was pointed out in Croux and Haesbroeck (2000), the M-estimator behaves particularly well in the Cauchy case but gives poor results for asymmetric contaminated distribution such as ACN2. The S-estimator is a high breakdown point estimator and works well even for asymmetric contaminated distribution. The estimate we propose has a

zero breakdown point but performs quite well on these simulated examples. When looking at the ρ estimation, it gives the worst results among the robust estimators for the Cauchy distribution but gives better results than the M-estimator for the ACN2 distribution. Note, however, that all the four estimators give poor results when estimating a nearly singular eigenvalue ($\lambda_6 = 0.02$) in presence of asymmetric contamination (ACN1 and ACN2).

22.4.2 Eigenvectors Estimators

Principal components are obtained by projection of the data on the eigenvectors estimates $\hat{v}_i$, $i = 1, \ldots, 6$. So, we measure the closeness between the estimator $\hat{v}_i$ and v_i by the absolute value of the cosine of the angle $\hat{\theta}_i$ between $\hat{v}_i$ and v_i. In Figs. 22.1, 22.2, 22.3, 22.4, 22.5, 22.6, we display the empirical cumulative distribution functions (ECDF) for the 1000 Monte Carlo realizations of $|\cos(\hat{\theta}_i)|$, $i = 1, 4, 6$, for the Cauchy and the ACN2 distributions which correspond to the most heavily contaminated distributions. Concerning the gaussian distribution, the figures are not reported here but the four empirical cumulative distribution functions curves are on top of each other. Concerning the SCN and the ACN1 distributions, the three robust estimators give similar results while the non-robust estimator does not perform very well. For the Cauchy and the ACN2 distributions, the cumulative

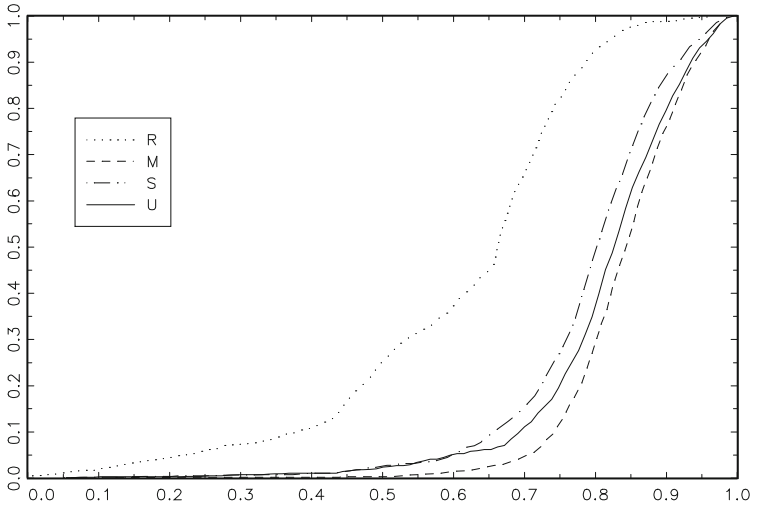

Fig. 22.1 ECDFs of $|\cos(\hat{\theta}_1)|$ at the CAU distribution for *R*, *M*, *S*, and *U*

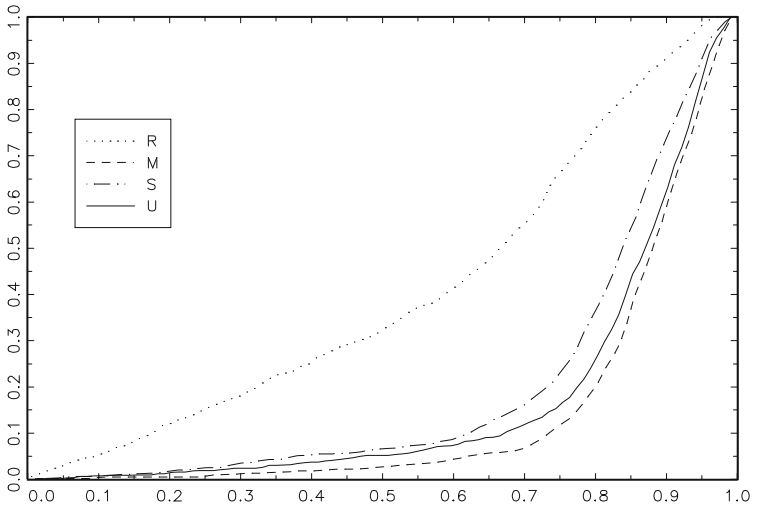

Fig. 22.2 ECDFs of $|\cos(\hat{\theta}_4)|$ at the CAU distribution for R, M, S, and U

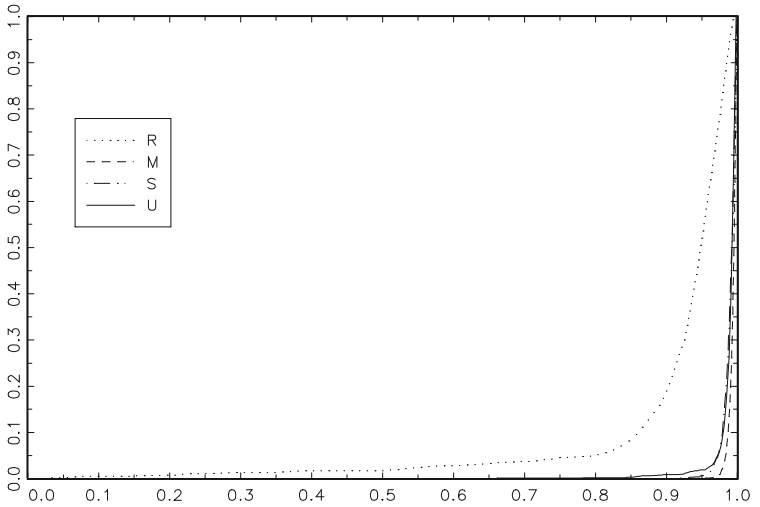

Fig. 22.3 ECDFs of $|\cos(\hat{\theta}_6)|$ at the CAU distribution for R, M, S, and U

functions of $|\cos(\hat{\theta}_i)|$, $i = 2, 3, 5$, are omitted since $|\cos(\hat{\theta}_6)|$ behaves like $|\cos(\hat{\theta}_5)|$ and $|\cos(\hat{\theta}_2)|$ and $|\cos(\hat{\theta}_3)|$ behave like $|\cos(\hat{\theta}_4)|$. The values of the $|\cos(\hat{\theta}_i)|$ should be close to one, so we can verify that the robust estimators outperform the non-robust one. The estimator U we propose behaves more or less like the M-estimator. The only case where M and U lack robustness is when estimating the direction of v_6

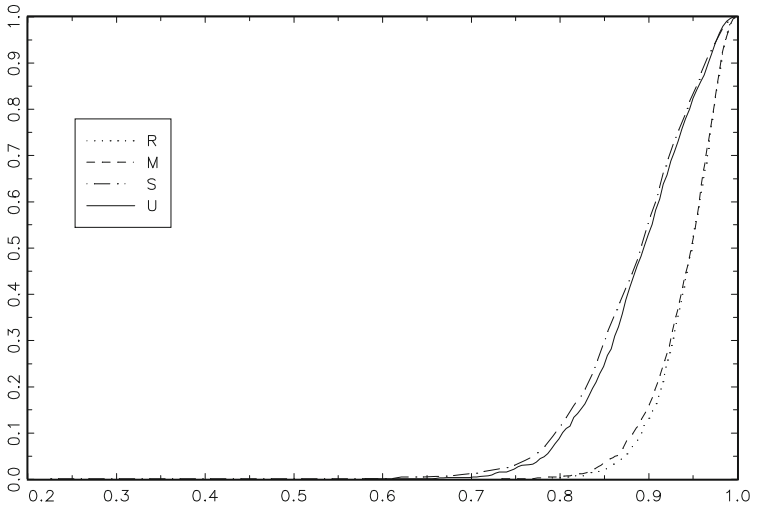

Fig. 22.4 ECDFs of $|\cos(\hat{\theta}_1)|$ at the ACN2 distribution for R, M, S, and U

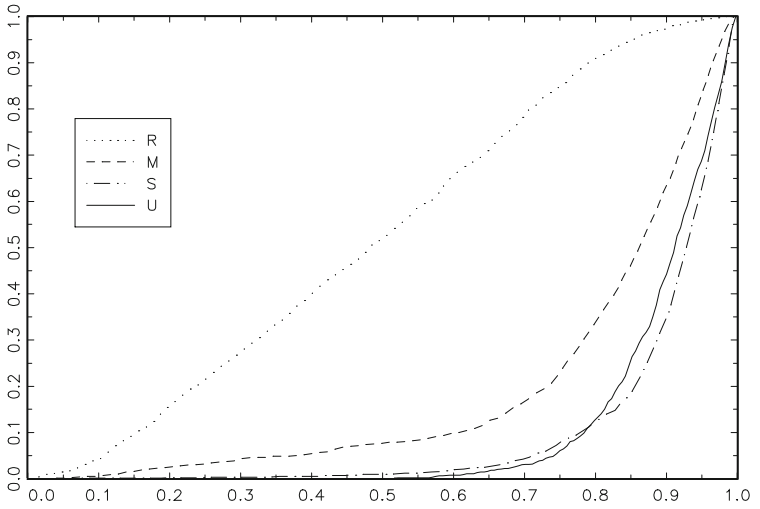

Fig. 22.5 ECDFs of $|\cos(\hat{\theta}_4)|$ at the ACN2 distribution for R, M, S, and U

in the ACN2 model (Fig. 22.6). More precisely, in this case, the estimated direction defined by the M and the U estimators is orthogonal to the direction defined by v_6. But this point is not really crucial if we are interested in Principal Component Analysis and focus on the largest eigenvalues in order to summarize the information contained in a large data set.

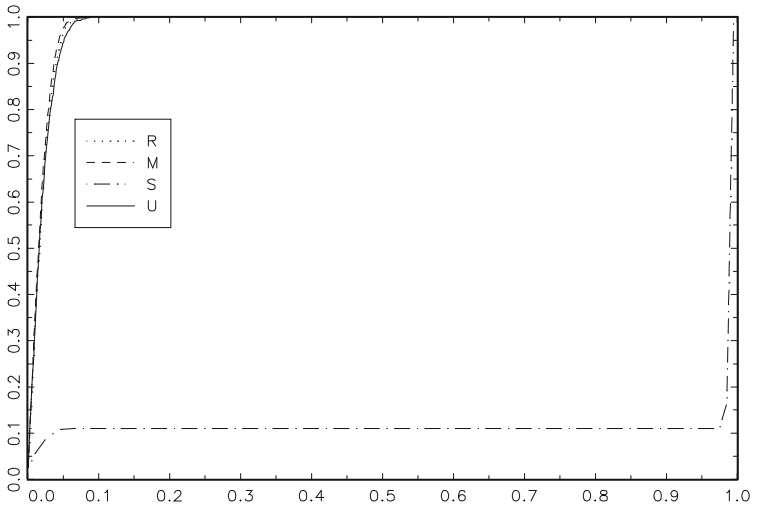

Fig. 22.6 ECDFs of $|\cos(\hat{\theta}_6)|$ at the ACN2 distribution for R, M, S, and U

22.5 Small Application

The main advantage of $\mathbf{U}_n(\beta)$ is that its calculation is simple and straightforward. Its drawback appears in situations where there is an important proportion of clustered outliers because of its zero breakdown point. In this section, we give an example of application to cluster detection using a particular ICS method where $\mathbf{U}_n(\beta)$ is advisable while high breakdown point estimators are not. Let us consider the invariant coordinate selection method as defined in Caussinus et al. (2003) with the objective of displaying groups of observations (if there are any in the data set). The method consists in an eigenvalue-eigenvector decomposition of one scatter estimator relative to another. In order to detect clusters, Caussinus et al. (2003) propose to use the usual covariance estimator together with a scatter matrix based on weighted pairwise differences. This pairwise-based estimator is estimating in some sense the within-cluster covariance matrix and the method can be interpreted as an unsupervised discriminant analysis. As it is often the case for exploratory projection pursuit methods, the interesting projections, namely those which display the groups, may be perturbed by a few extreme or atypical observations. In order to download the influence of these observations, we propose to use $\mathbf{U}_n(\beta)$ in place of the usual covariance matrix estimator. As it is illustrated in the following example, the use of high breakdown point estimators is clearly not indicated in this case since the method would not display the interesting projections anymore. The example we consider is the well known Lubischew data set (Lubischew 1962) which consists of $n = 74$ insects and $p = 6$ morphological measurements.

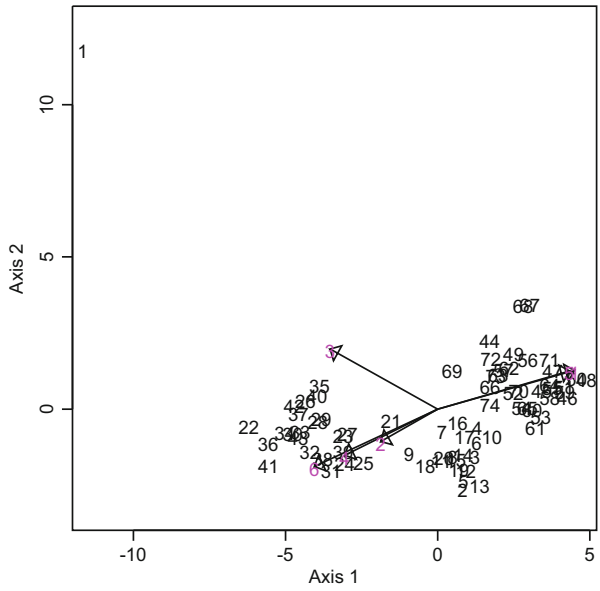

Fig. 22.7 First principal plane of invariant coordinate selection with the empirical non-robust dispersion matrix

According to taxonomy considerations, these insects are divided into eight groups. More precisely, observations 1 to 21 belong to the first group, observations 22 to 43 belong to the second while observations 44 to 74 belong to the third group. In order to illustrate the use of the robust covariance estimator $\mathbf{U}_n(\beta)$, we have changed the value of one of the variables for the first observation into an extreme value. Figure 22.7 gives the first projection plane obtained via the analysis when the non-robust empirical covariance estimator $\mathbf{V}_n$ is used while Fig. 22.8 gives the first principal plane when $\mathbf{V}_n$ is replaced by $\mathbf{U}_n(.2)$. The choice of $\beta = 0.2$ follows the recommendation given in Sect. 22.2.1 and is between $1/p \simeq 0.17$ and $(26\varepsilon)/p, \simeq 0.21$ for $\varepsilon = 5\,\%$. Only Fig. 22.8 clearly displays the three-group structure of the data. Indeed, the influence of the first outlying observation is dominant in Fig. 22.7 while the first observation is no more influential on the robust analysis. Figure 22.9 now gives the display obtained when $\mathbf{V}_n$ is replaced by a 25 % breakdown point S-estimator. This estimator is so robust that it does not only decrease the influence of the first observation. In some sense, the S-estimator is quite close to the within-cluster estimator but it is more robust to the first observation. In the context of ICS, this leads to the detection of the first observation as an outlier but not to the detection of groups. In this context, high breakdown point estimators are not recommended.

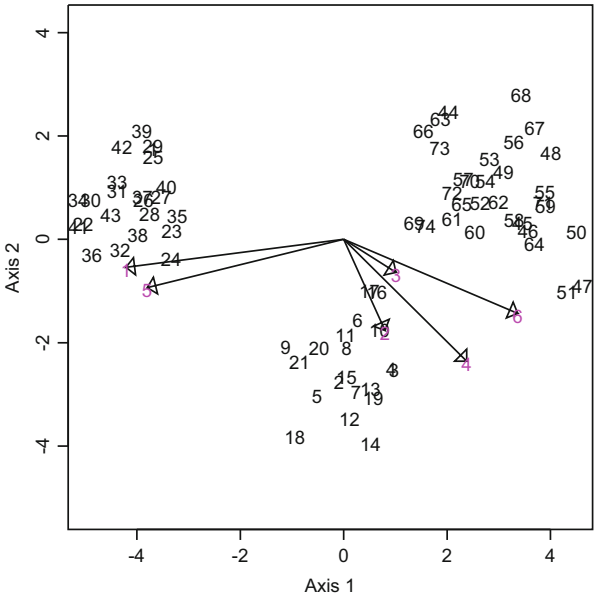

Fig. 22.8 First principal plane of invariant coordinate selection with the robust dispersion estimator $\mathbf{U}_n(.2)$

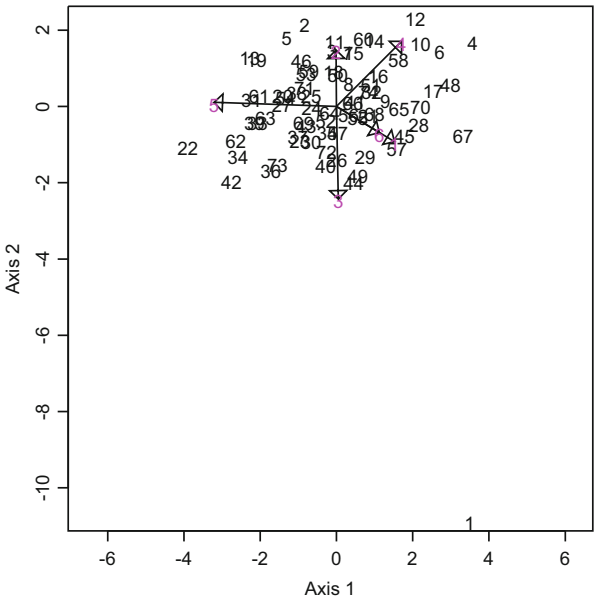

Fig. 22.9 First principal plane of invariant coordinate selection with a high breakdown point (25 %) S dispersion estimator

22.6 Conclusion

In this paper, the consistency and the asymptotic distribution of $\mathbf{U}_n(\beta)$ have been obtained using a uniform strong law of large number (Le Cam 1953) and some matrix manipulation. The eigenelements of a covariance or a correlation matrix are of great interest in multivariate data analysis such as Principal Component Analysis. Thus asymptotic distributions and influence functions are derived for the eigenelements of $\mathbf{U}_n(\beta)$ by using the perturbation theory (Kato 1966). The proposed estimator does not lose a lot of efficiency at the gaussian model and is robust from an infinitesimal point of view since the influence functions are bounded. Besides these theoretical properties, we also compare $\mathbf{U}_n(\beta)$ to other robust estimators by using the same simulation scheme as Devlin et al. (1981) and by performing an ICS analysis on a real data set. In spite of its zero breakdown point, $\mathbf{U}_n(\beta)$ performs well and is to be compared to M-estimators. $\mathbf{U}_n(\beta)$ can be calculated for large data sets, in the data mining context for instance. In fact, we recommend its systematic use in parallel with the usual covariance matrix estimator. If the two estimators give different results, since $\mathbf{U}_n(\beta)$ has infinitesimal robustness properties, it means that outlying observations are present in the data. Then, if heavily contaminated data are suspected, high-breakdown estimators such as the S- or the RMCD estimators are to be considered.

References

Bedall, F.K., Zimmerman, H.: Algorithm AS 143, the mediancenter. App. Stat. **28**, 325–328 (1979)

Béguin, C., Hulliger, B.: Multivariate outlier detection in incomplete survey data: the epidemic algorithm and transformed rank correlations. J. R. Stat. Soc. Ser. A (Stat. Soc.) **167**(2), 275–294 (2004)

Cator, E.A., Lopuhaä, H.P.: Central limit theorem and influence function for the MCD estimators at general multivariate distributions. Bernoulli **18**(2), 520–551 (2012)

Caussinus, H., Hakam, S., Ruiz-Gazen, A.: Projections révélatrices contrôlées: groupements et structures diverses. Rev. Stat. Appl. **51**(1), 37–58 (2003)

Critchley, F.: Influence in principal components analysis. Biometrika **72**, 627–636 (1985)

Croux, C., Haesbroeck, G.: Principal component analysis based on robust estimators of the covariance or correlation matrix : influence functions and efficiencies. Biometrika **87**(3), 603–618 (2000)

Dauxois, J., Pousse, A., Romain, Y.: Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. J. Multivar. Anal. **12**, 136–154 (1982)

Davies, P.L.: Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. Ann. Stat. **15**(3), 1269–1292 (1987)

Davies, L.: The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. Ann. Stat. **20**, 1828–1843 (1992)

Devlin, S.J., Gnanadesikan, R., Kettenring, J.R.: Robust estimation of dispersion matrices and principal components. J. Am. Stat. Assoc. **76**, 354–362 (1981)

Donoho, D.L.: Breakdown properties of multivariate location estimators. Ph.D. Qualifying Paper. Department of Statistics, Harvard University (1982)

Dossou-Gbete, S., Pousse, A.: Asymptotic study of eigenelements of a sequence of random selfadjoint operators. Statistics **3**, 479–491 (1991)
Dümbgen, L., Pauly, M., Schweizer, T.: A survey of M-functionals of multivariate location and scatter. arXiv preprint arXiv:1312.5594 (2013a)
Dümbgen, L., Nordhausen, K., Schuhmacher, H.: New algorithms for M-estimation of multivariate location and scatter. arXiv preprint arXiv:1312.6489 (2013b)
Fekri, M., Fine, J.: Matrice aléatoire dont l'espérance est de rang réduit. Propriétés asymptotiques des estimateurs moindres carrés et choix de métriques. Pub. Inst. Stat. Univ. Paris **XXXIX**(1), 67–88 (1995)
Fekri, M., Ruiz-Gazen, A.: Propriétés asymptotiques et fonction d'influence d'un estimateur simple et robuste de matrice de dispersion. C. R. Acad. Sci. Paris **t. 330** série I, 565–568 (2000)
Fischer, D., Möttönen, J., Nordhausen, K., Vogel, H.: OjaNP: multivariate methods based on the Oja median and related concepts. R package vesion 0.9-6. http://cran.r-project.org/web/packages/OjaNP/index.html (2013)
Gervini, D.: The influence function of the Stahel-Donoho estimator of multivariate location and scatter. Stat. Probab. Lett. **60**(4), 425–435 (2002)
Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.: Robust Statistics. Wiley, New York (1986)
Kato, T.: Perturbation Theory for Linear Operators. Springer, New York (1966)
Le Cam, L.: On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. Univ. Calif. Public Stat. **1**, 277–330 (1953)
Lopuhaä, H.P.: Asymptotics of reweighted estimators of multivariate location and scatter. Ann. Stat. **27**(5), 1638–1665 (1999)
Lubischew, A.A.: On the use of discriminant functions in taxonomy. Biometrics **18**(4), 455–477 (1962)
Ma, Y., Genton, M.G.: Highly robust estimation of dispersion matrices. J. Multiv. Anal. **78**(1), 11–36 (2001)
Maronna, R.A.: Robust M-estimators of multivariate location and scatter. Ann. Stat. **4**, 51–67 (1976)
Maronna, R.A., Stahel, W.A., Yohai, V.J.: Bias-robust estimators of multivariate scatter based on projections. J. Multiv. Anal. **42**(1), 141–161 (1992)
Maronna, R.A., Yohai, V.J.: Robust estimation of multivariate location and scatter. In: Kotz, S., Read, C., Banks, D. (eds.) Encyclopedia of Statistical Sciences, Update vol. 2, pp. 589–596. Wiley, New York (1988)
Maronna, R.A., Yohai, V.J.: The behavior of the Stahel-Donoho robust multivariate estimator. J. Am. Stat. Assoc. **90**, 330–341 (1995)
Maronna, R.A., Zamar, R.H.: Robust estimates of location and dispersion for high-dimensional datasets. Technometrics **44**(4), 307–317 (2002)
Meshalkin, L.D.: Approximation of multidimensional densities by normal distributions. In: Proceedings of 7th International Biometric Conference, Hannover (1970)
Oja, H.: Descriptive statistics for multivariate distributions. Stat. Probab. Lett. **1**(6), 327–332 (1983)
Ollila, E., Croux, C., Oja, H.: Influence function and asymptotic efficiency of the affine equivariant rank covariance matrix. DTEW Research Report 0210, 1–19 (2002)
Ollila, E., Oja, H., Croux, C.: The affine equivariant sign covariance matrix: asymptotic behavior and efficiencies. J. Multiv. Anal. **87**(2), 328–355 (2003)
Peña, D., Prieto, F.J.: Cluster identification using projections. J. Am. Stat. Assoc. **96**(456), 14433–1445 (2001)
Rousseeuw, P.J.: Multivariate estimation with high breakdown point. In: Grossmann, W., G. Pflug, G., I. Vincze, I., W. Wertz, W. (eds.) Mathematical Statistics and Applications, pp. 283–297. Reidel, Dordrecht (1985)
Rousseeuw, P.J., Croux, C.: The bias of k-step M-estimators. Stat. Probab. Lett. **20**(5), 411–420 (1994)
Rousseeuw, P.J., Leroy, A.M.: Robust Regression and Outlier Detection. Wiley, New York (1987)

Ruiz-Gazen, A.: A very simple robust estimator of a dispersion matrix. Comput. Stat. Data Anal. **21**, 149–162 (1996)

Stahel, W.A.: Breakdown of covariance estimators. Research Report 31. Fachgruppe für Statistik. ETH, Zürich (1981)

Tanaka, Y.: Sensitivity analysis in PCA: influence on the subspace spanned by principal components. Commun. Stat. Theory Methods **17**, 3157–3175 (1988)

Tyler, D.E., Critchley, F., Dümbgen, L., Oja, H.: Invariant co-ordinate selection. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **71**(3), 549–592 (2009)

Visuri, S., Koivunen, V., Oja, H.: Sign and rank covariance matrices. J. Stat. Plann. Inf. **91**(2), 557–575 (2000)

Zuo, Y., Cui, H., He, X.: On the Stahel–Donoho estimators and depth-weighted means for multivariate data. Ann. Stat. **32**(1), 167–188 (2004)

Zuo, Y., Lai, S.: Exact computation of bivariate projection depth and the Stahel–Donoho estimator. Comput. Stat. Data Anal. **55**(3), 1173–1179 (2011)

Chapter 23
On ANOVA-Like Matrix Decompositions

Giuseppe Bove, Frank Critchley, Radka Sabolova, and Germain Van Bever

Abstract The analysis of variance plays a fundamental role in statistical theory and practice, the standard Euclidean geometric form being particularly well established. The geometry and associated linear algebra underlying such standard analysis of variance methods permit, essentially direct, generalisation to other settings. Specifically, as jointly developed here: (a) to minimum distance estimation problems associated with subsets of pairwise orthogonal subspaces; (b) to matrix, rather than vector, contexts; and (c) to general, not just standard Euclidean, inner products, and their induced distance functions. To make such generalisation, we solve the following problem: given a set of nontrivial subspaces of a linear space, any two of which meet only at its origin, exactly which inner products make these subspaces pairwise *orthogonal*? Applications in a variety of areas are highlighted, including: (i) the analysis of asymmetry, and (ii) asymptotic comparisons in Invariant Coordinate Selection and Independent Component Analysis. A variety of possible further generalisations and applications are noted.

Keywords Analysis of asymmetry • Independent Components • Inner products • Invariant Coordinates • Orthogonal decomposition • Skew symmetry

23.1 Introduction

Geometry is a rich resource for statistical theory and practice. In particular, minimum distance estimation is an insightful, recurring, theme across many methodologies. Incorporating associated linear algebra, the analysis of variance (hereafter, ANOVA) is perhaps the stand-out example, estimation consisting essentially of orthogonal projection onto each of a set of pairwise orthogonal subspaces of a standard Euclidean space, whose direct sum is the space itself.

G. Bove
Dipartimento di Scienze della Formazione, Università degli Studi Roma Tre, Rome, Italy
e-mail: giuseppe.bove@uniroma3.it

F. Critchley (✉) • R. Sabolova • G. Van Bever
Department of Mathematics and Statistics, The Open University, Buckinghamshire, UK
e-mail: F.Critchley@open.ac.uk; radka.sabolova@open.ac.uk; germain.van-bever@open.ac.uk

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_23

Specifically, for some given dimension d, standard ANOVA takes place within $\mathscr{E}^d = (\mathscr{R}^d, \langle\cdot,\cdot\rangle_{\mathbf{I}})$—that is, within $\mathscr{R}^d$ endowed with the inner product $\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbf{I}} := \mathbf{x}^\top\mathbf{y}$, $(\mathbf{x}, \mathbf{y} \in \mathscr{R}^d)$, inducing the squared Euclidean distance between $\mathbf{x} = (x_i)$ and $\mathbf{y} = (y_i)$:

$$\|\mathbf{x} - \mathbf{y}\|_{\mathbf{I}}^2 = \textstyle\sum_{i=1}^d (x_i - y_i)^2.$$

Further, $\mathscr{E}^d$ is decomposed as the direct sum of $2 \le r \le d$ subspaces $\{\mathscr{L}_h\}_{h=1}^r$, any two of which are orthogonal. Accordingly, each $\mathbf{y} \in \mathscr{E}^d$ has unique decomposition

$$\mathbf{y} = \textstyle\sum_{h=1}^r \mathbf{y}_h \tag{23.1}$$

in which, for each $1 \le h \le r$, $\mathbf{y}_h$ is the orthogonal projection of $\mathbf{y}$ onto $\mathscr{L}_h$—that is, the nearest point in $\mathscr{L}_h$ to $\mathbf{y}$. Crucially, pairwise orthogonality of the subspaces means that *separate* minimum distance estimation of $\mathbf{y}$ by each $\mathscr{L}_h$ gives $\mathbf{y}_h$ and, thereby, the *overall* decomposition (23.1).

Essentially the same ideas can be applied more generally. First, and of central interest in this paper, let $\mathscr{L}_h' \subseteq \mathscr{L}_h$ $(h = 1, \ldots, r)$ be any given nonempty subsets of the original subspaces and suppose that, for given $\mathbf{y} \in \mathscr{E}^d$, we have the problem:

$$\text{minimise } \left\|\mathbf{y} - \textstyle\sum_{h=1}^r \hat{\mathbf{y}}_h\right\|_{\mathbf{I}}^2 \text{ subject to } \hat{\mathbf{y}}_h \in \mathscr{L}_h' \ (h = 1, \ldots, r). \tag{23.2}$$

Then, again, pairwise orthogonality of the $\{\mathscr{L}_h\}$ entails that separate minimum distance estimation of each $\mathbf{y}_h$ by $\mathscr{L}_h'$ gives $\hat{\mathbf{y}}_h$, these together solving the overall problem (23.2).

Secondly, vectors can be replaced by matrices. Section 23.2 reviews a variety of matrix analogues of the ANOVA decomposition (23.1). Each of these is based on the standard Euclidean inner product $\langle \mathbf{A}, \mathbf{B}\rangle_E := \text{trace}(\mathbf{A}^\top\mathbf{B})$. The induced (Frobenius) norm $\|\cdot\|_E$ has squared distance $\|\mathbf{A} - \mathbf{B}\|_E^2 = \sum_{i,j}(a_{ij} - b_{ij})^2$, the element-wise sum of squared differences. Applications in a variety of areas are highlighted, including: (i) the analysis of asymmetry (Constantine and Gower 1978; Gower 2014), and (ii) asymptotic comparisons in Invariant Coordinate Selection (ICS) (Tyler et al. 2009) and Independent Component Analysis (ICA) (Oja et al. 2006).

Thirdly, more general inner products can be used. In vector terms, the standard Euclidean inner product $\langle\cdot,\cdot\rangle_{\mathbf{I}}$ can be replaced by a general one:

$$\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbf{V}} := \mathbf{x}^\top\mathbf{V}\mathbf{y}$$

where $\mathbf{V}$ is any positive definite symmetric matrix of the appropriate order. Alternative inner products can also be used in a matrix context. For completeness' sake, Sect. 23.3 briefly reviews basic (essentially, standard) theory for these. Section 23.4 revisits analysis of asymmetry examples (Sect. 23.2.1) in this more general context. A variety of potential applications are noted.

The focus is on square matrices throughout. Extension to rectangular matrices, higher order arrays, and other vector space contexts is in principle straightforward.

To aid the flow of the paper, general terminology and definitions are introduced as we proceed. Its well-known or straightforward results are stated without proof.

23.2 Matrix Decomposition Examples

We briefly review here a variety of matrix decompositions based on standard Euclidean geometry. Sections 23.2.1 and 23.2.2 focus, respectively, on applications in the analysis of asymmetry and in ICS.

23.2.1 The Analysis of Asymmetry

Constantine and Gower (1978) emphasised how important, in practice, it can be to model systematic departures from symmetry in square data matrices. This influential paper gave rise to a field described here as the analysis of asymmetry, the recent review paper by Gower (2014) providing interesting historical perspective. To quote from its opening section, the driving motivation for this work was that:

> In practice, many applications are concerned just as much with any asymmetry as they are with symmetry. Well-known examples are studies in social mobility between classes, international trade between countries and pecking order in hens.

Additional technical details in Gower (2014) include, in particular, a formal proof of the form of the singular value decomposition of a skew-symmetric matrix.

Section 23.2.1.1 reviews a basic orthogonal decomposition of square matrices into their symmetric and skew-symmetric parts. In some contexts, a further decomposition may be called for. Section 23.2.1.2 treats a leading example of this. Section 23.2.1.3 discusses (i) a general orthogonal decomposition and (ii) other types of matrix subspace. Section 23.2.1.4 describes general benefits arising from using such decompositions, together with a worked example.

23.2.1.1 A Basic Decomposition

With the obvious definitions of addition and scalar multiplication of matrices, the set $\mathscr{M}_k$ of real square matrices of order k is a real linear (vector) space of dimension k^2. Omitting the subscript k, its subsets of symmetric and skew(-symmetric) members, $\mathscr{S} := \{\mathbf{S} \in \mathscr{M} : \mathbf{S}^\top = \mathbf{S}\}$ and $\mathscr{K} := \{\mathbf{K} \in \mathscr{M} : \mathbf{K}^\top = -\mathbf{K}\}$, are *subspaces*—i.e. are closed under addition and scalar multiplication—with dimensions $\binom{k+1}{2}$ and $\binom{k}{2}$, respectively. Moreover,

$$\mathscr{S} + \mathscr{K} := \{\mathbf{S} + \mathbf{K} : \mathbf{S} \in \mathscr{S}, \mathbf{K} \in \mathscr{K}\} = \mathscr{M}, \text{ while } \mathscr{S} \cap \mathscr{K} = \{\mathbf{0}_{\mathscr{M}}\}. \tag{23.3}$$

That is (definition), $\mathscr{S}$ and $\mathscr{K}$ have *direct sum* $\mathscr{M}$, denoted $\mathscr{M} = \mathscr{S} \oplus \mathscr{K}$. Equivalently, every $\mathbf{M} \in \mathscr{M}$ can be *uniquely* represented in the form $\mathbf{M} = \mathbf{S}_\mathbf{M} + \mathbf{K}_\mathbf{M}$ where $\mathbf{S}_\mathbf{M} \in \mathscr{S}$ and $\mathbf{K}_\mathbf{M} \in \mathscr{K}$. Explicitly:

$$\mathbf{S}_\mathbf{M} = \tfrac{1}{2}(\mathbf{M} + \mathbf{M}^\top) \text{ and } \mathbf{K}_\mathbf{M} = \tfrac{1}{2}(\mathbf{M} - \mathbf{M}^\top). \tag{23.4}$$

If we endow $\mathscr{M}$ with the standard Euclidean inner product, reviewed in the Introduction, the direct sum relations (23.3) sharpen to:

Proposition 23.1 *$\mathscr{M} = \mathscr{S} \oplus \mathscr{K}$ in which $\mathscr{S} = \mathscr{K}^{\perp}$ and $\mathscr{K} = \mathscr{S}^{\perp}$ are orthogonal complements in the inner product space $(\mathscr{M}, \langle \cdot, \cdot \rangle_E)$. Accordingly, for every $\mathbf{M} \in \mathscr{M}$,*

$$\|\mathbf{M}\|_E^2 = \|\mathbf{S}_\mathbf{M}\|_E^2 + \|\mathbf{K}_\mathbf{M}\|_E^2 . \tag{23.5}$$

In a seminal paper, Constantine and Gower (1978) exploited this result to advocate separate fitting of models to the symmetric and skew parts of a given data matrix $\mathbf{M}$. For, if $\mathscr{S}' \subseteq \mathscr{S}$ and $\mathscr{K}' \subseteq \mathscr{K}$ denote any symmetric and skew matrix model classes, minimisation of

$$\|\mathbf{M} - (\mathbf{S} + \mathbf{K})\|_E^2 = \|\mathbf{S}_\mathbf{M} - \mathbf{S}\|_E^2 + \|\mathbf{K}_\mathbf{M} - \mathbf{K}\|_E^2$$

over all $(\mathbf{S}, \mathbf{K}) \in \mathscr{S}' \times \mathscr{K}'$ is, palpably, accomplished by *separate* least squares fitting of $\mathbf{S}_\mathbf{M}$ by $\mathbf{S} \in \mathscr{S}'$ and of $\mathbf{K}_\mathbf{M}$ by $\mathbf{K} \in \mathscr{K}'$.

23.2.1.2 A Further Decomposition

Consider the following decomposition of $\mathscr{S}$. Denoting by $*$ the Hadamard (direct) product of matrices,

$$\mathscr{D} := \{\mathbf{D} \in \mathscr{S} : \mathbf{D} * \mathbf{I} = \mathbf{D}\} \text{ and } \mathscr{H} := \{\mathbf{H} \in \mathscr{S} : \mathbf{H} * \mathbf{I} = \mathbf{0}_\mathscr{M}\}$$

comprise all diagonal and all "hollow" (zero diagonal) members of $\mathscr{S}$, these being subspaces of $\mathscr{S}$ of dimension k and $\binom{k}{2}$, respectively, and we have:

$$\mathscr{S} = \mathscr{D} \oplus \mathscr{H},$$

$\mathbf{S} \in \mathscr{S}$ being uniquely expressible as $\mathbf{S} = \mathbf{D}_\mathbf{S} + \mathbf{H}_\mathbf{S}$ with $\mathbf{D}_\mathbf{S} \in \mathscr{D}$ and $\mathbf{H}_\mathbf{S} \in \mathscr{H}$. Explicitly, $\mathbf{D}_\mathbf{S} = \mathbf{S} * \mathbf{I}$ and $\mathbf{H}_\mathbf{S} = \mathbf{S} * (\mathbf{1}\mathbf{1}^\top - \mathbf{I})$, where $\mathbf{1}$ denotes the vector of ones.

Overall, $\mathscr{M} = (\mathscr{D} \oplus \mathscr{H}) \oplus \mathscr{K}$ which, the order of summation being clearly irrelevant, we may write as $\mathscr{M} = \mathscr{D} \oplus \mathscr{H} \oplus \mathscr{K}$, each $\mathbf{M} \in \mathscr{M}$ having unique decomposition in terms of these subspaces as:

$$\mathbf{M} = \mathbf{D}_{\mathbf{S}_\mathbf{M}} + \mathbf{H}_{\mathbf{S}_\mathbf{M}} + \mathbf{K}_\mathbf{M}. \tag{23.6}$$

If we endow $\mathscr{M}$ with the standard Euclidean inner product, inherited by each of its subspaces, the above direct sum relations sharpen to:

Proposition 23.2 *$\mathscr{M} = \mathscr{D} \oplus \mathscr{H} \oplus \mathscr{K}$ in which $\mathscr{D}$, $\mathscr{H}$ and $\mathscr{K}$ are pairwise orthogonal subspaces of the inner product space $(\mathscr{M}, \langle \cdot, \cdot \rangle_E)$. Accordingly, for every*

$\mathbf{M} \in \mathscr{M}$,

$$\|\mathbf{M}\|_E^2 = \|\mathbf{D}_{\mathbf{S_M}}\|_E^2 + \|\mathbf{H}_{\mathbf{S_M}}\|_E^2 + \|\mathbf{K}_{\mathbf{M}}\|_E^2.$$

Again, this result justifies separate fitting of models to the diagonal, hollow and skew parts of a given data matrix $\mathbf{M}$. For, if $\mathscr{D}' \subseteq \mathscr{D}$, $\mathscr{H}' \subseteq \mathscr{H}$ and $\mathscr{K}' \subseteq \mathscr{K}$ denote any diagonal, hollow and skew matrix model classes, minimisation of

$$\|\mathbf{M} - (\mathbf{D} + \mathbf{H} + \mathbf{K})\|_E^2 = \|\mathbf{D}_{\mathbf{S_M}} - \mathbf{D}\|_E^2 + \|\mathbf{H}_{\mathbf{S_M}} - \mathbf{H}\|_E^2 + \|\mathbf{K}_{\mathbf{M}} - \mathbf{K}\|_E^2$$

over all $(\mathbf{D}, \mathbf{H}, \mathbf{K}) \in \mathscr{D}' \times \mathscr{H}' \times \mathscr{K}'$ is clearly accomplished by separate least squares fitting of each of $(\mathbf{D}_{\mathbf{S_M}}, \mathbf{H}_{\mathbf{S_M}}, \mathbf{K}_{\mathbf{M}})$ within its model class.

23.2.1.3 A General Decomposition and Other Subspace Types

The methodology developed above generalises directly to decomposition of $\mathscr{M}$ into any collection of any number of pairwise orthogonal subspaces in an obvious way, analogous—indeed, formally, isomorphic—to familiar orthogonal decompositions of $\mathscr{E}^d$ used in standard ANOVA.

Replacing subspaces of vectors in $\mathscr{E}^d$ by subspaces of matrices in $\mathscr{M}$, we have

Proposition 23.3 *Let $\{\mathscr{L}_h\}_{h=1}^r$ $(2 \le r \le k^2)$ be a collection of nontrivial subspaces of $\mathscr{M}$, any two of which are orthogonal in $(\mathscr{M}, \langle\cdot,\cdot\rangle_E)$, and whose direct sum $\oplus\{\mathscr{L}_h\}$ is $\mathscr{M}$. Then, every $\mathbf{M} \in \mathscr{M}$ can be uniquely expressed as $\mathbf{M} = \sum_{h=1}^r \mathbf{M}_h$, in which $\mathbf{M}_h$ is the orthogonal projection of $\mathbf{M}$ onto $\mathscr{L}_h$, so that*

$$\|\mathbf{M}\|_E^2 = \textstyle\sum_{h=1}^r \|\mathbf{M}_h\|_E^2.$$

As a direct corollary, for any given model classes $\mathscr{L}_h' \subseteq \mathscr{L}_h$ $(h = 1, \ldots, r)$ within the "parent" subspaces $\{\mathscr{L}_h\}$, and for any given $\mathbf{M} \in \mathscr{M}_k$, the overall problem:

$$\text{minimise } \left\|\mathbf{M} - \textstyle\sum_{h=1}^r \hat{\mathbf{M}}_h\right\|_E^2 \text{ subject to } \hat{\mathbf{M}}_h \in \mathscr{L}_h' \ (h = 1, \ldots, r) \tag{23.7}$$

can be solved by separate minimum distance estimation of each $\mathbf{M}_h$ by $\mathscr{L}_h'$.

There is a wide variety of types of subspace of potential interest in modelling data taking the form of a square matrix. In particular, the subspaces $\mathscr{L}_{\mathscr{Z}}$ of matrices (m_{ij}) for which m_{ij} vanishes for each $(i,j) \in \mathscr{Z}$, $\emptyset \subseteq \mathscr{Z} \subseteq (1, \ldots, k)^2$, the orthogonal complement $\mathscr{L}_{\mathscr{Z}}^{\perp}$ of each such subspace being given by:

$$\mathscr{L}_{\mathscr{Z}}^{\perp} = \mathscr{L}_{\mathscr{Z}^C}.$$

Subspaces of this type include matrices vanishing on a given set of rows or columns.

Other instances of $\mathscr{L}_{\mathscr{Z}}$ are of special interest when the rows and columns of $\mathbf{M}$ have the same natural order. For example—in a context where m_{ij} is modelled as

vanishing when i and j belong to different members of a partition of $(1,\ldots,k)$ into consecutive blocks—the corresponding blockdiagonal matrices. Or—in a context where $(1,\ldots,k)$ label consecutive time points and m_{ij} is held to vanish for i and j sufficiently far apart in time—the (say) tri-diagonal matrices defined by: $|i-j| > 1 \Rightarrow m_{ij} = 0$.

Or, again, in social mobility studies—with $(1,\ldots,k)$ labelling a classification of households from high to low social status, and m_{ij} the number of adults in a category i household whose parents formed a category j household—the strictly upper triangular, diagonal and strictly lower triangular matrices are of interest. Specifically, with $\mathscr{L}_{\uparrow}$, $\mathscr{L}_{\rightarrow}$ and $\mathscr{L}_{\downarrow}$ denoting the subspaces $\mathscr{L}_{\mathscr{Z}}$ corresponding, respectively, to $\mathscr{Z} = \mathscr{Z}_{\uparrow} := \{(i,j) \in (1,\ldots,k)^2 : j \leq i\}$, $\mathscr{Z} = \mathscr{Z}_{\rightarrow} := \{(i,j) \in (1,\ldots,k)^2 : i \neq j\}$ and $\mathscr{Z} = \mathscr{Z}_{\downarrow} := \{(i,j) \in (1,\ldots,k)^2 : i \leq j\}$, we have the orthogonal decomposition:

$$\mathscr{M} = \mathscr{L}_{\uparrow} \oplus \mathscr{L}_{\rightarrow} \oplus \mathscr{L}_{\downarrow}$$

in which $\mathscr{L}_{\uparrow}$ and $\mathscr{L}_{\downarrow}$ correspond to upward and downward inter-generational social mobility, respectively, while $\mathscr{L}_{\rightarrow}$ corresponds to absence of either form of mobility. Typically, different models will be appropriate to these three processes, so that the expected value of a data matrix $\mathbf{M}$ is best modelled as the sum of matrices from each of three corresponding—often, parameterised—model classes $\mathscr{L}'_{\uparrow}$, $\mathscr{L}'_{\rightarrow}$ and $\mathscr{L}'_{\downarrow}$.

More generally than $\mathscr{L}_{\mathscr{Z}}$, the symmetric and skew-symmetric subspaces $\mathscr{S}$ and $\mathscr{K}$ are examples of subspaces on which a prescribed set of linear relations hold among the elements of $\mathbf{M}$.

Again, the set of subspaces of $\mathscr{M}$ is closed under orthogonal complementation, intersection and addition, giving "new subspaces for old". For example, $\mathscr{L}_{\rightarrow}^{\perp}$ comprises all matrices with zero diagonal entries, while $\mathscr{H} = \mathscr{L}_{\rightarrow}^{\perp} \cap \mathscr{S}$. Again, the upper and lower triangular matrices form the subspaces $\mathscr{L}_{\uparrow} \oplus \mathscr{L}_{\rightarrow}$ and $\mathscr{L}_{\downarrow} \oplus \mathscr{L}_{\rightarrow}$, respectively.

Finally in this section, we note general benefits flowing from use of orthogonal decompositions of $\mathscr{M}$ as $\oplus\{\mathscr{L}_h\}$, as in Proposition 23.3, together with a worked example using the decomposition $\mathscr{M} = \mathscr{D} \oplus \mathscr{H} \oplus \mathscr{K}$.

23.2.1.4 General Benefits and a Worked Example

As illustrated by the social mobility example above, the expected value of a data matrix $\mathbf{M} \in \mathscr{M}$ can often be best modelled as a sum of r matrices ($2 \leq r \leq k^2$), one from each of a set of model classes $\{\mathscr{L}'_h\}_{h=1}^{r}$ where $\{\mathbf{0}_{\mathscr{M}}\} \subset \mathscr{L}'_h \subseteq \mathscr{L}_h$ ($1 \leq h \leq r$) and $\oplus_{h=1}^{r}\{\mathscr{L}_h\}$ is an orthogonal decomposition of $(\mathscr{M}, \langle\cdot,\cdot\rangle_E)$ into component subspaces. Assuming uncorrelated homoscedastic additive errors with zero mean and common variance σ^2, least-squares fitting—equivalently, assuming also Gaussianity, maximum likelihood estimation—then leads to the general constrained matrix approximation problem (23.7).

Exploiting orthogonality, this problem can be solved by separate minimum distance estimation of each $\mathbf{M}_h$ by $\mathscr{L}'_h$. This breaking up of an overall problem into component parts can be especially beneficial, computationally, when the model classes $\{\mathscr{L}'_h\}$ depend in complex nonlinear ways on separate underlying parameter vectors $\{\boldsymbol{\theta}_h\}$.

Again, such separate fitting breaks up underlying variability into its component parts. In an obvious notation, we have

$$\sigma^2 = \sum_{h=1}^{r} \sigma_h^2.$$

Each of the components of variance σ_h^2 is estimable, in the obvious way. Indeed, under Gaussianity, the resulting $\{\hat{\sigma}_h^2\}$ are mutually independent, with the usual benefits for inference.

Finally, borrowing a term from generalised linear models, undefined or non-stochastic elements of a data matrix $\mathbf{M}$ can be accommodated by introducing corresponding *offset* terms.

To fix ideas, consider the problem of modelling a matrix $\mathbf{M}$ of scheduled flight times between cities in a part of the world subject to a strong prevailing wind blowing from West to East (say). A natural approach, described below, uses the orthogonal decomposition $\mathscr{M} = \mathscr{D} \oplus \mathscr{H} \oplus \mathscr{K}$.

The diagonal entries of $\mathbf{M}$ being undefined, we don't try to model them. Rather, formally, we set $\mathbf{D}_{\mathbf{S_M}} = \mathbf{0}_{\mathscr{M}}$, taking the corresponding model class as $\mathscr{D}' = \{\mathbf{0}_{\mathscr{M}}\}$. These offset terms assure perfect diagonal "fit", effectively replacing minimisation of $||\mathbf{M} - \hat{\mathbf{M}}||_E^2$ with that of

$$||\mathbf{M} - \hat{\mathbf{M}}||_{E'}^2 := \sum_{i \neq j} (m_{ij} - \hat{m}_{ij})^2.$$

Consider now the off-diagonal entries of $\mathbf{M}$. By definition, $\mathbf{H}_{\mathbf{S_M}}$ belongs to $\mathscr{H}^+ := \{(h_{ij}) \in \mathscr{H} : \text{each } h_{ij} \geq 0\}$, the set of all $k \times k$ dissimilarity matrices, and so can be fitted by a suitable multidimensional scaling method. Alternatively, as developed here, we may assume that the expected value of the average of the scheduled flight time from city i to city j and that for the opposite journey is a function of g_{ij}, the geodesic distance between them. As one parametric example, we may take $\mathscr{H}' = \mathscr{H}'(\boldsymbol{\theta}_{\mathbf{H}})$ in which $\boldsymbol{\theta}_{\mathbf{H}} = (\alpha, \beta, \kappa)^\top$, the general member $\mathbf{H}$ of $\mathscr{H}'(\boldsymbol{\theta}_{\mathbf{H}})$ having (i,j)th element $\alpha + \beta g_{ij}^{\kappa}$. The constant term α here reflects a fixed time needed for take off and landing.

We turn now to the skew symmetric part of $\mathbf{M}$. Constantine and Gower (1978) take an essentially exploratory approach, based on the singular value decomposition of $\mathbf{K}_{\mathbf{M}}$ and, in particular, graphical representation of its dominant parts. Alternatively, for each $i < j$, the expected mean difference k_{ij} between the scheduled flight time from city i to city j and that for the opposite journey can be taken as an odd function of $(a_i - a_j)$, where a_i denotes the longitude of city i. As one parametric example, we may take the general member $\mathbf{K}$ of $\mathscr{K}' = \mathscr{K}'(\boldsymbol{\theta}_{\mathbf{K}})$ as having (i,j)th element $k_{ij} = \gamma(a_i - a_j)$, so that $\theta_{\mathbf{K}} = \gamma$. Introducing additional parameters,

essentially the same model can be elaborated to include estimation of an *unknown* prevailing wind direction.

Separate minimum distance estimation of $\mathbf{H}_{\mathbf{S_M}}$ by $\mathscr{H}'$, and of $\mathbf{K}_{\mathbf{M}}$ by $\mathscr{K}'$, speeds computation of the optimal estimates of $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathbf{H}}^{\top}, \boldsymbol{\theta}_{\mathbf{K}}^{\top})^{\top}$. This computational gain is, in general, greater the larger the number of components being fitted and the more complicated the parametric functional forms involved.

Again, this separation into sub-problems allows (estimation of) the decomposition $\sigma^2 = \sigma_{\mathbf{H}}^2 + \sigma_{\mathbf{K}}^2$. Under Gaussianity, $\hat{\boldsymbol{\theta}}_{\mathbf{H}}$ and $\hat{\boldsymbol{\theta}}_{\mathbf{K}}$ are independent. Accordingly, so too are $\hat{\sigma}_{\mathbf{H}}^2$ and $\hat{\sigma}_{\mathbf{K}}^2$.

23.2.2 Independent Component Analysis and Invariant Coordinate Selection

A somewhat different use of the $\langle \cdot, \cdot \rangle_E$-orthogonal decomposition $\mathscr{M} = \mathscr{S} \oplus \mathscr{K}$ arises in connection with ICA—see, for example, Oja et al. (2006); and with ICS—see Tyler et al. (2009).

ICA is a highly popular method within many applied areas. Its principal objective is to recover, as far as is possible, unobserved independent components from an observable random vector arising as an unknown (possibly shifted) convolution of them. Oja et al. (2006) were the first to point out that, under appropriate modelling assumptions, this recovery—or "unmixing"—problem can be solved via what is now known as ICS.

ICS is a general method for exploring multivariate data, based on two scatter matrices. Scatter matrices—generalisations of the usual covariance operator—are matrix-valued functionals $\mathbf{S} = \mathbf{S}(F_{\mathbf{X}})$ of the distribution of a random vector $\mathbf{X}$ that are affine equivariant. That is, for any nonsingular matrix $\mathbf{C}$ and any vector $\mathbf{c}$, $\mathbf{X} \to \mathbf{X}^* := \mathbf{C}\mathbf{X} + \mathbf{c}$ induces $\mathbf{S} \to \mathbf{S}^* = \mathbf{C}\mathbf{S}\mathbf{C}^{\top}$. One of these matrices, denoted $\mathbf{V}$ below, is required to be positive definite symmetric.

ICS works by transforming $\mathbf{X} \to \mathbf{Z} = \mathbf{M}^{\top}\mathbf{X}$, where $\mathbf{M} = \mathbf{U}\mathbf{V}^{-1/2}$ with $\mathbf{U}$ an orthogonal matrix determined jointly by the two scatter matrices. The elements of $\mathbf{Z}$ are called "invariant coordinates" to connote the fact that any nonsingular affine transformation $\mathbf{X} \to \mathbf{X}^*$ leaves $\mathbf{Z}$ unchanged (up to a shift term that vanishes if and only if $\mathbf{c} = 0$).

As Oja showed, for suitable ICA models, we have the striking result that *the invariant coordinates correspond to the independent components*. The question then naturally arises as to whether a pair of scatter matrices $(\mathbf{V}, \tilde{\mathbf{V}})$ can be chosen that is, in some sense, optimal in this regard. Here, fixing $\mathbf{V}$ as the usual covariance operator, we note the following. Ilmonen et al. (2010) establish that, in the standard case,

$$\sqrt{n}(\mathbf{V} - \mathbf{I}) = O_P(1) \text{ and } \sqrt{n}(\mathbf{U} - \mathbf{I}) = O_P(1). \tag{23.8}$$

Recent correspondence with that paper's final author, Hannu Oja, notes that (23.8) implies

$$\sqrt{n}(\mathbf{S_M}-\mathbf{I}) = -\tfrac{1}{2}\sqrt{n}(\mathbf{V}-\mathbf{I}) + o_P(1) \text{ and } \sqrt{n}\mathbf{K_M} = \sqrt{n}(\mathbf{U}-\mathbf{I}) + o_P(1). \quad (23.9)$$

This may have implications for comparing two estimates obtained with the same $\mathbf{V}$. Relations (23.9) suggest that asymptotic comparison of estimates obtained with fixed $\mathbf{V}$ but different $\tilde{\mathbf{V}}$—hence, different $\mathbf{U}$—should be based on the upper triangular part of $\mathbf{U}$ only (and not on the full information contained in $\mathbf{M}$). In contrast, currently, such comparison is usually made based on the off-diagonal elements of $\mathbf{M}$, introducing an unnecessary confounding with its symmetric part.

Overall, it will be of interest to explore potential methodological benefits flowing from the direct sum $\mathscr{M} = \mathscr{S} \oplus \mathscr{K}$ underpinning (23.9). In this connection, note also that the decomposition $\mathscr{M} = \mathscr{D} \oplus \mathscr{H} \oplus \mathscr{K}$ can be used to refine (23.9). And that the sharpening by orthogonality of these direct sum relations in Propositions 23.1 and 23.2, and their generalisations to other inner products established in Sect. 23.4.2 below, may bring additional benefits. In particular, it may be possible to helpfully link the choice of inner product to the asymptotic precision (inverse covariance) matrices of the estimators involved.

23.3 Alternative Inner Products

Orthogonality—in the standard Euclidean inner product $\langle \cdot, \cdot \rangle_E$—is central to all the key results above and, thus, to their sphere of applicability. At the same time, the (often, implicit) assumption of uncorrelated homoscedastic errors being made is, as always, open to question.

Theoretical inquisitiveness combines, then, with a desire for the widest possible realm of practical applications to point to the importance of finding answers to the following, natural, questions: for what *other* inner products do Propositions 23.1 and 23.2 hold? More generally (*cf.* Proposition 23.3), given a collection $\{\mathscr{L}_h\}$ of nontrivial subspaces with direct sum $\mathscr{M}$, any two of which are orthogonal in $(\mathscr{M}, \langle \cdot, \cdot \rangle_E)$, for what *other* inner products does such pairwise orthogonality hold? (Specific proposals for such alternative inner products, in a variety of contexts, are discussed in Sect. 23.4.3.)

In order to answer these questions, specific to the vector (sub)spaces introduced in Sect. 23.2, it will be sufficient—and actually straightforward—to answer the following generic question. Let $\mathscr{L}$ be a real linear (vector) space with finite dimension $d \geq 2$. Let now $\{\mathscr{L}_h\}_{h=1}^r$ $(2 \leq r \leq d)$ be a collection of nontrivial subspaces of $\mathscr{L}$ whose direct sum, $\oplus\{\mathscr{L}_h\}$, is $\mathscr{L}$. That is,

$$\mathscr{L}_h \cap \mathscr{L}_{h'} = \{\mathbf{0}_{\mathscr{L}}\}, (h \neq h'), \text{ while } \mathscr{L} = \textstyle\sum_h \mathscr{L}_h := \{\mathbf{l}_1 + \ldots + \mathbf{l}_m : \mathbf{l}_h \in \mathscr{L}_h\}, \quad (23.10)$$

(so that $d = \sum_h d_h$, where $\mathscr{L}_h$ has dimension $d_h > 0$).

The question posed is: in which inner product spaces $(\mathscr{L}, \langle\cdot,\cdot\rangle)$ are the $\{\mathscr{L}_h\}$ pairwise orthogonal? That is, for which inner products does $\langle \mathbf{l}_h, \mathbf{l}_{h'}\rangle$ vanish for each $\mathbf{l}_h \in \mathscr{L}_h$ and $\mathbf{l}_{h'} \in \mathscr{L}_{h'}$, $(h \neq h')$? As will be clear, the answer consists, essentially, in taking *ordered* bases.

Recall that, if $\mathscr{L}^*$ is also a real linear space, $T : \mathscr{L} \to \mathscr{L}^*$ is called an isomorphism if it is one-to-one, onto and linear: that is, for all α_1, α_2 in $\mathscr{R}$ and all $\mathbf{l}_1, \mathbf{l}_2$ in $\mathscr{L}$,

$$T(\alpha_1 \mathbf{l}_1 + \alpha_2 \mathbf{l}_2) = \alpha_1 T(\mathbf{l}_1) + \alpha_2 T(\mathbf{l}_2).$$

$\mathscr{L}$ and $\mathscr{L}^*$ are called isomorphic if such a map exists, and we write $\mathscr{L} \cong \mathscr{L}^*$ (via T).

Proposition 23.4 *Let $\{\mathbf{b}_1, \ldots, \mathbf{b}_d\}$ be a basis for $\mathscr{L}$. Then, $\mathscr{L} \cong \mathscr{R}^d$ via $\mathbf{l} \to \boldsymbol{\xi}$ where, to each $\mathbf{l} = \sum_i \xi_i \mathbf{b}_i \in \mathscr{L}$, we associate $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_d)^\top \in \mathscr{R}^d$.*

Recall also that an inner product $\langle\cdot,\cdot\rangle$ on $\mathscr{L}$ is a function from $\mathscr{L}^2$ to $\mathscr{R}$ such that for all $\mathbf{l}, \mathbf{l}', \mathbf{l}_1, \mathbf{l}_2$ in $\mathscr{L}$ and α_1, α_2 in $\mathscr{R}$:

1. $\langle \mathbf{l}, \mathbf{l}'\rangle = \langle \mathbf{l}', \mathbf{l}\rangle$
2. $\langle \alpha_1 \mathbf{l}_1 + \alpha_2 \mathbf{l}_2, \mathbf{l}'\rangle = \alpha_1 \langle \mathbf{l}_1, \mathbf{l}'\rangle + \alpha_2 \langle \mathbf{l}_2, \mathbf{l}'\rangle$
3. $\langle \mathbf{l}, \mathbf{l}\rangle \geq 0$, equality holding if and only if $\mathbf{l} = \mathbf{0}_{\mathscr{L}}$.

Being bilinear, an inner product is completely determined by its matrix of values $(\langle \mathbf{b}_i, \mathbf{b}_j\rangle)$ on an ordered basis $(\mathbf{b}_1, \ldots, \mathbf{b}_d)$. Accordingly, denoting by $\mathscr{V}_d$ the set of all real, positive definite symmetric matrices of order d, the following result is immediate.

Proposition 23.5 *For any ordered basis $\mathbf{b}(\mathscr{L}) = (\mathbf{b}_1, \ldots, \mathbf{b}_d)$ for $\mathscr{L}$, $\langle\cdot,\cdot\rangle \to (\langle \mathbf{b}_i, \mathbf{b}_j\rangle)$ is a one-to-one correspondence between the set of all inner products on $\mathscr{L}$ and $\mathscr{V}_d$.*

Recall further that an inner product isomorphism between $(\mathscr{L}, \langle\cdot,\cdot\rangle)$ and $(\mathscr{L}^*, \langle\cdot,\cdot\rangle^*)$ is an isomorphism $T : \mathscr{L} \to \mathscr{L}^*$ which preserves inner products: that is, such that for all $\mathbf{l}_1, \mathbf{l}_2$ in $\mathscr{L}$: $\langle T(\mathbf{l}_1), T(\mathbf{l}_2)\rangle^* = \langle \mathbf{l}_1, \mathbf{l}_2\rangle$. $(\mathscr{L}, \langle\cdot,\cdot\rangle)$ and $(\mathscr{L}^*, \langle\cdot,\cdot\rangle^*)$ are called isomorphic if such a map exists, and we write $(\mathscr{L}, \langle\cdot,\cdot\rangle) \cong (\mathscr{L}^*, \langle\cdot,\cdot\rangle^*)$ (via T). Combining Propositions 23.4 and 23.5, we have:

Proposition 23.6 *Let $\mathbf{b}(\mathscr{L}) = (\mathbf{b}_1, \ldots, \mathbf{b}_d)$ be an ordered basis for $\mathscr{L}$, let $\langle\cdot,\cdot\rangle$ be an inner product on $\mathscr{L}$, and let $\mathbf{V} = (\langle \mathbf{b}_i, \mathbf{b}_j\rangle)$. Then $(\mathscr{L}, \langle\cdot,\cdot\rangle) \cong (\mathscr{R}^d, \langle\cdot,\cdot\rangle_{\mathbf{V}})$ via $T_{\mathbf{b}(\mathscr{L})} : \mathbf{l} \to \boldsymbol{\xi}$, where $\langle \boldsymbol{\xi}, \boldsymbol{\eta}\rangle_{\mathbf{V}} := \boldsymbol{\xi}^\top \mathbf{V} \boldsymbol{\eta}$.*

Denoting by $\mathscr{Q}_d$ (respectively: $\mathscr{D}_d^+$) the set of all orthogonal (respectively: diagonal, positive definite) matrices of order d, and by $diag(\cdot, \ldots, \cdot)$ the blockdiagonal matrix with the diagonal entries listed, the simple answer to the generic question posed is:

Theorem 23.1 *Let $\mathscr{L}$ and $\{\mathscr{L}_h\}_{h=1}^r$ $(2 \leq r \leq d)$ be as in (23.10). Let $(\mathbf{b}_1, \ldots, \mathbf{b}_d)$ be an ordered basis for $\mathscr{L}$ comprising, successively, ordered bases for each of*

$\mathscr{L}_1, \ldots, \mathscr{L}_r$. *Then, the following are equivalent:*

1. *the* $\{\mathscr{L}_h\}$ *are pairwise orthogonal in* $(\mathscr{L}, \langle \cdot, \cdot \rangle)$,
2. $(\langle \mathbf{b}_i, \mathbf{b}_j \rangle) = diag(\mathbf{V}_1, \ldots, \mathbf{V}_r)$ *for some* $\mathbf{V}_h \in \mathscr{V}_{d_h}$, $h = 1, \ldots, r$,
3. $(\langle \mathbf{b}_i, \mathbf{b}_j \rangle) = \mathbf{Q}\boldsymbol{\Delta}\mathbf{Q}^T$ *for some* $\mathbf{Q} = diag(\mathbf{Q}_1, \ldots, \mathbf{Q}_r)$ *and* $\boldsymbol{\Delta} = diag(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_r)$, *where* $\mathbf{Q}_h \in \mathscr{Q}_{d_h}$ *and* $\boldsymbol{\Delta}_h \in \mathscr{D}^+_{d_h}$, $h = 1, \ldots, r$.

Theorem 23.1 is couched, of course, in terms of a chosen basis $(\mathbf{b}_1, \ldots, \mathbf{b}_d)$. To see what form $\mathbf{V} = (\langle \mathbf{b}_i, \mathbf{b}_j \rangle)$ takes in a general basis, it suffices to use the following. Bilinearity of inner products gives at once:

Proposition 23.7 *For any inner product* $\langle \cdot, \cdot \rangle$ *on* $\mathscr{L}$, *and for any ordered basis* $(\mathbf{b}_1, \ldots, \mathbf{b}_d)$ *of* $\mathscr{L}$, *let* $\mathbf{V} := (\langle \mathbf{b}_i, \mathbf{b}_j \rangle)$. *For any other ordered basis* $(\mathbf{b}^*_1, \ldots, \mathbf{b}^*_d)$ *of* $\mathscr{L}$,

$$\mathbf{V}^* := (\langle \mathbf{b}^*_i, \mathbf{b}^*_j \rangle) = \mathbf{A}\mathbf{V}\mathbf{A}^\top$$

where the nonsingular matrix $\mathbf{A} = (a_{ij})$ *is given by* $\mathbf{b}^*_i = \sum_j a_{ij}\mathbf{b}_j$.

23.4 Matrix Decomposition Examples Revisited

In the wider context of inner products preserving the same pairwise orthogonalities, we revisit here the analysis of asymmetry matrix decomposition examples reviewed in Sect. 23.2.1.

We first introduce (Sect. 23.4.1) two alternative ordered bases for $\mathscr{M}_k$, and then (Sect. 23.4.2) use them to generalise two key results underpinning the standard Euclidean analysis of asymmetry, viz.: Proposition 23.1 (Sect. 23.2.1.1) and Proposition 23.2 (Sect. 23.2.1.2). Section 23.4.3 indicates applications of these more general results.

23.4.1 Two Alternative Ordered Bases for $\mathscr{M}_k$

In this asymmetry context, rather than the usual stacking-by-columns operator, it is convenient to order the elements of $\mathbf{M} \in \mathscr{M}_k$ by first listing those on the diagonal; then, those above the diagonal, in row-wise order; and, finally, those below the diagonal, in column-wise order. In this way, we define the *veck* operator by

$$\begin{aligned} veck(\mathbf{M}) = &(m_{11}, \ldots, m_{kk}; \\ &m_{12}, \ldots, m_{1k}, m_{23}, \ldots, m_{k-1,k}; \\ &m_{21}, \ldots, m_{k1}, m_{32}, \ldots, m_{k,k-1})^\top. \end{aligned}$$

We use relevant parts of this order to define ordered bases of $\mathscr{D}$, $\mathscr{H}$, $\mathscr{S}$, $\mathscr{K}$ and $\mathscr{M} = \mathscr{M}_k$ itself in terms of the canonical basis $\{\mathbf{E}_{ij} := \mathbf{e}_i\mathbf{e}_j^\top, 1 \le i,j \le k\}$ of $\mathscr{M}$. Here, $(\mathbf{e}_1, \ldots, \mathbf{e}_k)$ is the usual ordered basis of $\mathscr{R}^k$, so that $\mathbf{e}_i$ is a binary vector with a single one in the ith position. For notational convenience, the constant ρ below is $1/\sqrt{2}$.

Putting $\mathbf{D}_{ii} := \mathbf{E}_{ii}$, $\mathbf{b}(\mathscr{D}) := (\mathbf{D}_{11}, \ldots, \mathbf{D}_{kk})$ is an ordered basis for $\mathscr{D}$. Again, with $\mathbf{H}_{ij} := \rho(\mathbf{E}_{ij} + \mathbf{E}_{ji})$, $i < j$, $\mathbf{b}(\mathscr{H}) := (\mathbf{H}_{12}, \ldots, \mathbf{H}_{k-1,k})$ is an ordered basis for $\mathscr{H}$. Concatenating these, in a mild abuse of notation, $\mathbf{b}(\mathscr{S}) := (\mathbf{b}(\mathscr{D}); \mathbf{b}(\mathscr{H}))$ is an ordered basis for $\mathscr{S}$. Finally, with $\mathbf{K}_{ij} := \rho(\mathbf{E}_{ij} - \mathbf{E}_{ji})$, $i < j$, $\mathbf{b}(\mathscr{K}) := (\mathbf{K}_{12}, \ldots, \mathbf{K}_{k-1,k})$ is an ordered basis for $\mathscr{K}$, so that $\mathbf{b}(\mathscr{M}) := (\mathbf{b}(\mathscr{S}); \mathbf{b}(\mathscr{K}))$ is an ordered basis for $\mathscr{M}$.

A natural alternative is to use the corresponding ordering of the canonical basis itself:

$$\mathbf{b}^*(\mathscr{M}) = (\mathbf{E}_{11}, \ldots, \mathbf{E}_{kk}; \mathbf{E}_{12}, \ldots, \mathbf{E}_{k-1,k}; \mathbf{E}_{21}, \ldots, \mathbf{E}_{k,k-1})$$

which corresponds directly to the data we actually observe, since:

$$\text{each } \mathbf{M} \in \mathscr{M}_k \text{ has direct decomposition } \mathbf{M} = \textstyle\sum_{i,j} m_{ij}\mathbf{E}_{ij}. \tag{23.11}$$

By design, $\mathbf{b}(\mathscr{M})$ and $\mathbf{b}^*(\mathscr{M})$ are ordered *orthonormal* bases in $(\mathscr{M}, \langle\cdot,\cdot\rangle_E)$. Since $\mathbf{E}_{ij} = \rho(\mathbf{H}_{ij} + \mathbf{K}_{ij})$ while $\mathbf{E}_{ji} = \rho(\mathbf{H}_{ij} - \mathbf{K}_{ij})$, $i < j$, the orthogonal matrix $\mathbf{A}$ linking these two bases via $\mathbf{b}_i^*(\mathscr{M}) = \sum_j a_{ij}\mathbf{b}_j(\mathscr{M})$ is given by:

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_k & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \rho\mathbf{I}_{\binom{k}{2}} & \rho\mathbf{I}_{\binom{k}{2}} \\ \mathbf{0} & \rho\mathbf{I}_{\binom{k}{2}} & -\rho\mathbf{I}_{\binom{k}{2}} \end{pmatrix} = diag(\mathbf{I}_k, \mathbf{B} \otimes \mathbf{I}_{\binom{k}{2}}), \tag{23.12}$$

$\otimes$ denoting the Kronecker product, so that $\mathbf{A}$ (hence, $\mathbf{B}$) is also symmetric, the matrix

$$\mathbf{B} = \rho\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \rho\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \cdot \rho\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

combining, for each $i < j$, a 45° rotation and a reflection in the two-dimensional subspace spanned by $(\mathbf{H}_{ij}, \mathbf{K}_{ij})$; equivalently, by $(\mathbf{E}_{ij}, \mathbf{E}_{ji})$. Note that, each being orthogonal and symmetric, $\mathbf{A}$ and $\mathbf{B}$ are self-inverse.

From now on, for any inner product $\langle\cdot,\cdot\rangle$ on $\mathscr{M}$, we define $\mathbf{V} \equiv \mathbf{V}_{(\mathscr{M},\langle\cdot,\cdot\rangle)}$ and $\mathbf{V}^* \equiv \mathbf{V}^*_{(\mathscr{M},\langle\cdot,\cdot\rangle)}$ thus:

$$\mathbf{V} := (\langle \mathbf{b}_i(\mathscr{M}), \mathbf{b}_j(\mathscr{M}) \rangle) \text{ and } \mathbf{V}^* := (\langle \mathbf{b}_i^*(\mathscr{M}), \mathbf{b}_j^*(\mathscr{M}) \rangle), \tag{23.13}$$

so that, by Proposition 23.7, $\mathbf{V}^* = \mathbf{A}\mathbf{V}\mathbf{A}^T$; equivalently, as $\mathbf{A}^{-1} = \mathbf{A}$, $\mathbf{V} = \mathbf{A}\mathbf{V}^*\mathbf{A}^T$.

Proposition 23.6 and display (23.11) give at once

Proposition 23.8 $\mathbf{M} \to \mathbf{m} := vec(\mathbf{M})$ *is an inner product isomorphism between* $(\mathscr{M}_k, \langle\cdot,\cdot\rangle)$ *and* $(\mathscr{R}^{k^2}, \langle\cdot,\cdot\rangle_{\mathbf{V}^*})$, *where* $\langle \mathbf{m}_1, \mathbf{m}_2\rangle_{\mathbf{V}^*} := \mathbf{m}_1^\top \mathbf{V}^* \mathbf{m}_2$.

In particular, $(\mathscr{M}_k, \langle\cdot,\cdot\rangle_E) \cong (\mathscr{R}^{k^2}, \langle\cdot,\cdot\rangle_{\mathbf{I}})$ *via* $\mathbf{M} \to vec(\mathbf{M})$.

The standard Euclidean inner product exploited in Propositions 23.1 and 23.2 corresponds, then, to taking $\mathbf{V}^* = \mathbf{I}$; equivalently, $\mathbf{V} = \mathbf{I}$. We show next exactly what other choices Theorem 23.1 allows.

23.4.2 Generalisations of Propositions 23.1 and 23.2

Recall the matrices $\mathbf{A}$, $\mathbf{V}$ and $\mathbf{V}^*$ defined in (23.12) and (23.13) above.

Considering first generalisations of Proposition 23.1, Theorem 23.1 gives at once that $\mathscr{S}$ and $\mathscr{K}$ are orthogonal in $(\mathscr{M}, \langle\cdot,\cdot\rangle)$ if and only if

$$\mathbf{V} = diag(\mathbf{V_S}, \mathbf{V_K}) \text{ for some } \mathbf{V_S} \in \mathscr{V}_{\binom{k+1}{2}} \text{ and } \mathbf{V_K} \in \mathscr{V}_{\binom{k}{2}}.$$

Partitioning any such $\mathbf{V}$ conformably with $\mathbf{A}$, we write it as

$$\mathbf{V} = \begin{pmatrix} \mathbf{V_D} & \mathbf{V_{DH}} & \mathbf{0} \\ \mathbf{V_{DH}}^\top & \mathbf{V_H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V_K} \end{pmatrix}. \tag{23.14}$$

Concerning generalisations of Proposition 23.2, Theorem 23.1 again gives at once that $\mathscr{D}$, $\mathscr{H}$ and $\mathscr{K}$ are pairwise orthogonal in $(\mathscr{M}, \langle\cdot,\cdot\rangle)$ if and only if $\mathbf{V}$, as in (23.14), has $\mathbf{V_{DH}} = \mathbf{0}$, so that $\mathbf{V_D} \in \mathscr{V}_k$ and $\mathbf{V_H} \in \mathscr{V}_{\binom{k}{2}}$. For general $\mathbf{V_{DH}}$–see, for example, Theorem 7.7.6 in Horn and Johnson (1985, p. 472)—we have that

Proposition 23.9 *The following are equivalent:*

1. $\mathbf{V_S} \equiv \begin{pmatrix} \mathbf{V_D} & \mathbf{V_{DH}} \\ \mathbf{V_{DH}}^\top & \mathbf{V_H} \end{pmatrix} \in \mathscr{V}_{\binom{k+1}{2}}$
2. $\mathbf{V_D} \in \mathscr{V}_k$ *and* $[\mathbf{V_H} - \mathbf{V_{DH}}^\top \mathbf{V_D}^{-1} \mathbf{V_{DH}}] \in \mathscr{V}_{\binom{k}{2}}$
3. $\mathbf{V_H} \in \mathscr{V}_{\binom{k}{2}}$ *and* $[\mathbf{V_D} - \mathbf{V_{DH}} \mathbf{V_H}^{-1} \mathbf{V_{DH}}^\top] \in \mathscr{V}_k$.

It only remains to find the form of $\mathbf{V}^*$. Using Propositions 23.7 and 23.9, and recalling that $\rho = 1/\sqrt{2}$, (23.12) gives

Theorem 23.2 $\mathscr{S}$ *and* $\mathscr{K}$ *are orthogonal in* $(\mathscr{M}, \langle\cdot,\cdot\rangle)$ *if and only if*

$$\mathbf{V}^* = \begin{pmatrix} \mathbf{V_D} & \rho\mathbf{V_{DH}} & \rho\mathbf{V_{DH}} \\ \rho\mathbf{V_{DH}}^\top & \rho^2(\mathbf{V_H} + \mathbf{V_K}) & \rho^2(\mathbf{V_H} - \mathbf{V_K}) \\ \rho\mathbf{V_{DH}}^\top & \rho^2(\mathbf{V_H} - \mathbf{V_K}) & \rho^2(\mathbf{V_H} + \mathbf{V_K}) \end{pmatrix} \tag{23.15}$$

for some $\mathbf{V_K} \in \mathcal{V}_{\binom{k}{2}}$, $\mathbf{V_H} \in \mathcal{V}_{\binom{k}{2}}$ *and* $\mathbf{V_{DH}}$ *as in (23.14) with* $[\mathbf{V_D} - \mathbf{V_{DH}}\mathbf{V_H^{-1}}\mathbf{V_{DH}^{\top}}] \in \mathcal{V}_k$, $\mathscr{D}$ *and* $\mathscr{H}$ *also being orthogonal if and only if* $\mathbf{V_{DH}} = \mathbf{0}$.

Theorem 23.2 has two immediate corollaries. By design, pairwise orthogonality in $(\mathscr{M}, \langle\cdot,\cdot\rangle)$ of the specified set of subspaces, $\{\mathscr{S}, \mathscr{K}\}$ or $\{\mathscr{D}, \mathscr{H}, \mathscr{K}\}$, is reflected in the corresponding blockdiagonal form of $\mathbf{V}$. The first corollary characterises when $\mathbf{V}^*$ enjoys the *same* form of blockdiagonal structure, showing indeed that this happens if and only if $\mathbf{V}^* = \mathbf{V}$.

Corollary 23.1 *Let* $\langle\cdot,\cdot\rangle$ *be any inner product on* $\mathscr{M}$ *for which* $\mathscr{S}$ *and* $\mathscr{K}$ *are orthogonal. Then, equivalent are:*

1. $\mathbf{V}^* = diag(\mathbf{V}_a, \mathbf{V}_b)$ *for some* $\mathbf{V}_a \in \mathcal{V}_{\binom{k+1}{2}}$ *and* $\mathbf{V}_b \in \mathcal{V}_{\binom{k}{2}}$
2. $\mathbf{V}^* = diag(\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3)$ *for some* $\mathbf{V}_1 \in \mathcal{V}_k$ *and* $\mathbf{V}_2, \mathbf{V}_3 \in \mathcal{V}_{\binom{k}{2}}$
3. $\mathbf{V}^* = diag(\mathbf{V_D}, \overline{\mathbf{V}}, \overline{\mathbf{V}})$ *for some* $\overline{\mathbf{V}} \in \mathcal{V}_{\binom{k}{2}}$
4. $\mathbf{V_{DH}} = \mathbf{0}$ *and* $\mathbf{V_H} = \mathbf{V_K}$
5. $\mathbf{V}^* = \mathbf{V}$.

The second, relatively trivial, corollary characterises when $\mathbf{V}^*$ is fully diagonal, so that the inner product on $\mathscr{M}$ corresponds to a set of weights: $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i,j} w_{ij} a_{ij} b_{ij}$ for some positive $\{w_{ij}\}$. We have that the only requirement for orthogonality of $\mathscr{S}$ and $\mathscr{K}$ is the symmetry condition $w_{ji} = w_{ij}$, $i < j$. That is:

Corollary 23.2 *Let* $\langle\cdot,\cdot\rangle$ *be any inner product on* $\mathscr{M}$ *for which* $\mathscr{S}$ *and* $\mathscr{K}$ *are orthogonal. Then, equivalent are:*

1. $\mathbf{V}^* \in \mathscr{D}^+_{k^2}$
2. $\mathbf{V}^* = diag(\boldsymbol{\Delta}_1, \overline{\boldsymbol{\Delta}}, \overline{\boldsymbol{\Delta}})$ *for some* $\boldsymbol{\Delta}_1 \in \mathscr{D}^+_k$ *and* $\overline{\boldsymbol{\Delta}} \in \mathscr{D}^+_{\binom{k}{2}}$.

23.4.3 Applications

Theorems 23.1 and 23.2 determine all possible choices of $\mathbf{V}$—equivalently, of $\mathbf{V}^*$—consistent with the orthogonalities required for use in generalising Proposition 23.1 or 23.2. In practice, this choice of inner product will depend on context. We briefly indicate three ways in which it may be made.

One way is to allow for alternative forms of the covariance of the observed data. Recall first the vector case, reviewed briefly in the Introduction. Here, standard ANOVA methods—those based on Euclidean distances, induced by $\langle ., . \rangle_{\mathbf{I}}$—are appropriate when $cov(\mathbf{y})$ is taken as proportional to $\mathbf{I}$. The move from ordinary to generalised least-squares ANOVA methods is signalled whenever, in contrast, error terms are taken as correlated and/or heteroscedastic. Effectively, in this case, Euclidean distances are replaced by their Mahalanobis counterparts, induced by $\langle ., . \rangle_{\mathbf{V}_y}$ where $\mathbf{V}_y = (cov(\mathbf{y}))^{-1}$ is assumed known up to a positive scalar multiple. Similarly, in the matrix case reviewed in Sect. 23.2.1.4, standard ANOVA

methods—those based on distances induced by $\langle ., . \rangle_E$—are appropriate for uncorrelated homoscedastic errors. However, a generally appropriate choice is to take $\mathbf{V}^*$, of the form given in Theorem 23.2, as a positive scalar multiple of the inverse of an assumed (asymptotic) covariance matrix for $vec(\mathbf{M})$. In particular, weighted least-squares estimation is appropriate in the uncorrelated but heteroscedastic case, when we may take $\mathbf{V}^*$ to have the diagonal form characterised in Corollary 23.2.

An alternative way of choosing an inner product focuses on $\mathbf{V}$. Recall that a nonsingular covariance matrix has blockdiagonal form if and only if the same is true of its inverse, the number and size of the blocks being the same in both cases. A second choice is then to model the blockdiagonal form of $\mathbf{V}$ required by orthogonality as reflecting uncorrelatedness of the corresponding parts of $\mathbf{M}$. Corollary 23.1 shows that this happens if and only if $\mathbf{V_{DH}} = \mathbf{O}$ and $\mathbf{V_H} = \mathbf{V_K}$. Equivalently, when $\mathbf{V}^* = \mathbf{V}$. This modelling choice is very much in the same "analyse separately" spirit as the seminal (Constantine and Gower 1978) paper. The within-subspace dispersions may then be modelled according to context.

Finally, under multivariate normality—and so, in many asymptotic situations—zeroes in precision (inverse covariance) matrices have particular significance in terms of conditional independence. Depending on context, such conditional independence considerations may further guide appropriate choice of $\mathbf{V}$ or $\mathbf{V}^*$. In particular, there may be useful applications here to graphical model contexts.

Acknowledgements The UK-based authors thank the EPSRC for their support under grant EP/L010429/1.

References

Constantine, A.G., Gower, J.C.: Graphical representation of asymmetric matrices. J. R. Stat. Soc. Ser. C **27**, 297–304 (1978)

Gower, J.C.: Skew symmetry in retrospect. Adv. Data Anal. Classif. (2014). doi: 10.1007/s11634-014-0181-7

Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1985)

Ilmonen, P., Nevalainen, J., Oja, H.: Characteristics of multivariate distributions and the invariant coordinate system. Stat. Probab. Lett. **80**, 1844–1853 (2010)

Oja, H., Sirkiä, S., Eriksson, J.: Scatter matrices and independent component analysis. Austrian J. Stat. **35**, 175–189 (2006)

Tyler, D.E., Critchley, F., Dümbgen, L., Oja, H.: Invariant coordinate selection (with discussion). J. R. Stat. Soc. Ser. B **71**, 549–592 (2009)

Chapter 24
On Invariant Within Equivalence Coordinate System (IWECS) Transformations

Robert Serfling

Abstract In exploratory data analysis and data mining in the very common setting of a data set $\mathbb{X}$ of vectors from $\mathbb{R}^d$, the search for important features and artifacts of a geometrical nature is a leading focus. Here one must insist that such discoveries be invariant under selected changes of coordinates, at least within some specified equivalence relation on geometric structures. Otherwise, interesting findings could be merely artifacts of the coordinate system. To avoid such pitfalls, it is desirable to transform the data $\mathbb{X}$ to an associated data cloud $\mathbb{X}^*$ whose geometric structure may be viewed as intrinsic to the given data $\mathbb{X}$ but also invariant in the desired sense. General treatments of such "invariant coordinate system" transformations have been developed from various perspectives. As a timely step, here we formulate a more structured and unifying framework for the relevant concepts. With this in hand, we develop results that clarify the roles of the so-called transformation-retransformation transformations. We illustrate by treating invariance properties of some outlyingness functions. Finally, we examine productive connections with maximal invariants.

Keywords Invariant coordinate systems • Maximal invariants • Multivariate • Nonparametric • Outlyingness functions • Transformations

24.1 Introduction

In exploratory data analysis and data mining in the very common setting of a data set $\mathbb{X}$ of vectors from $\mathbb{R}^d$, a leading focus is the search for important features and artifacts of a geometrical nature. Here one should insist that such discoveries be invariant under selected changes of coordinates, or at least be invariant under such changes up to a particular equivalence relation on geometric structures. Otherwise, what appears to be interesting geometric structure could be nothing but an artifact of the particular coordinate system adopted. To avoid such a pitfall, the data $\mathbb{X}$ can be transformed to an associated new data cloud $\mathbb{X}^*$ having geometric structure that is

R. Serfling (✉)
Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75083, USA
e-mail: serfling@utdallas.edu

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_24

intrinsically related to $\mathbb{X}$ but also invariant in the desired sense. General treatments of such "invariant coordinate system" transformations have been developed from various perspectives. As a timely next step, here we introduce a more structured and unifying framework for the relevant concepts, develop results that clarify the use of transformation-retransformation transformations, illustrate by treating invariance of some popular outlyingness functions, and productively examine connections with maximal invariants.

The topic of transformation to an "*invariant coordinate system (ICS)*" is treated broadly in the seminal paper of Tyler et al. (2009), and further general treatments are provided in Ilmonen et al. (2010), Serfling (2010), and Ilmonen et al. (2012). See also Nordhausen (2008) for useful results. Collectively, these sources treat two quite different approaches toward construction of "ICS" transformations and discuss a diversity of interesting practical applications.

However, the various treatments to date are not completely coherent and precise with respect to what is actually meant by "ICS." Indeed, for many of the examples and applications, the desired invariance is achieved only within some equivalence relation defined on the geometric structures of data sets. For example, for a data set $\mathbb{X}$, when we seek to identify its geometric structure that is invariant under affine transformation, it might be the case for the given application that differences due to homogeneous scale changes, coordinatewise sign changes, and translations may be ignored. That is, for any affine transformation of the given data cloud $\mathbb{X}$ to $\mathbb{Y}$, we might require only that the corresponding invariant coordinate systems $\mathbb{X}^*$ and $\mathbb{Y}^*$ agree only within such a specified equivalence. To accommodate a variety of such practical applications, the notions and terminology of "ICS" have evolved very productively but in somewhat loose fashion.

It is now timely and useful to have a more structured conceptual framework that draws together the various "ICS" results and adds perspective. For this purpose, we introduce and study in Sect. 24.2 a precise notion of "*invariant within equivalence coordinate system (IWECS)*" transformation: $\mathbf{M}(\mathbb{X})$ such that the transformed data $\mathbf{M}(\mathbb{X})\mathbb{X}$ is invariant under transformation of $\mathbb{X}$ relative to a transformation group $\mathscr{G}$, subject to equivalence relative to another transformation group $\mathscr{F}$. That is, for $g \in \mathscr{G}$, $\mathbf{M}(\mathbb{X})\mathbb{X}$ and $\mathbf{M}(g\mathbb{X})g\mathbb{X}$ need not be equal but must fall in the same equivalence class. Specifically, $\mathbf{M}(\mathbb{X})\mathbb{X}$ is to be $\mathscr{G}$-invariant within $\mathscr{F}$-equivalence.

It is seen in Serfling (2010) that the ICS transformations of practical interest fall within the class of transformation-retransformation (TR) transformations, which are essentially inverse square roots of covariance matrices. The chief purpose of TR transformations is standardization of data, so that estimators, test statistics, and other sample statistics become affine invariant or equivariant when defined on the standardized data. However, in some such cases a strong type of TR transformation is needed, namely an ICS transformation. Also, it is of interest to know when a TR transformation may directly play the role of an ICS transformation. Here we note that, to serve additionally as an ICS transformation, a TR matrix must be rather atypical, since the "usual" ICS transformations cannot be symmetric or triangular (Serfling 2010). In Sect. 24.3 we provide further clarifications on TR versus ICS transformations, as follows. Theorem 24.2 provides the narrowest equivalence (i.e., the smallest $\mathscr{F}$) for which TR transformations can serve as IWECS transformations

relative to linear or affine invariance. Theorem 24.3 exhibits a key special class of TR transformations for which the corresponding IWECS are affine invariant relative to the smallest possible nontrivial choice of $\mathscr{F}$. As illustrations of the application of these theorems, we treat affine invariant TR versions of the spatial outlyingness function and of the projection outlyingness function when the number of projections used is finite.

The construction of TR matrices that possess the structural properties requisite to be ICS (or IWECS) is somewhat challenging. Relative to the linear and affine transformation groups, connections between a useful strong special case of ICS and IWECS transformations and the relevant maximal invariant statistics are examined in Sect. 24.4.1. Thus maximal invariant statistics can play a role in constructing ICS and IWECS transformations in this special case. We provide background references on two distinctive approaches that have been developed along these lines. Further exploiting connections with maximal invariant statistics, in Sect. 24.4.2 we revisit classical treatments (Lehmann 1959) of maximal invariants relative to these groups and "discover" a competitive third approach, one offering greater simplicity and less computational burden.

The present paper treats only the case of data from a Euclidean space, as does all of the literature to date except for extension to complex-valued data (Ilmonen 2013). However, the concepts we present in fact can have very general extension and potentially have application in quite diverse contexts. We provide brief discussion in Sect. 24.5.

As the literature we cite in this paper amply portrays, there has been a prominent guiding influence in developing, studying, and applying ICS transformations. The contributions of the present paper are dedicated as a tribute to Hannu Oja and his leadership.

24.2 A General Framework for Formulation of Invariant Within Equivalence Coordinate Systems in $\mathbb{R}^d$

24.2.1 General Framework

Here we draw together and extend recent general treatments of ICS transformations (Nordhausen 2008; Tyler et al. 2009; Ilmonen et al. 2010; Serfling 2010; Ilmonen et al. 2012). A general framework for describing the inherent geometrical structure of a data set in $\mathbb{R}^d$ via *IWECS representations* is defined as follows.

Definition 24.1 An *IWECS* in $\mathbb{R}^d$ consists of three components $(\mathscr{G}, \mathbf{M}(\cdot), \mathscr{E})$,

1. a group $\mathscr{G}$ of transformations g on data sets $\mathbb{X}$ of observations from $\mathbb{R}^d$,
2. a data-based $d \times d$ matrix transformation $\mathbf{M}(\mathbb{X})$ taking $\mathbb{X}$ to $\mathbf{M}(\mathbb{X})\mathbb{X}$, and
3. an equivalence relation $\mathscr{E}$ on the transformed data sets $\mathbf{M}(\mathbb{X})\mathbb{X}$,

such that $\mathbf{M}(g\mathbb{X})g\mathbb{X}$, $g \in \mathscr{G}$, all lie in the same equivalence class relative to $\mathscr{E}$, i.e., are invariant relative to $\mathscr{G}$ within $\mathscr{E}$-equivalence. □

Thus the $\mathscr{E}$-orbit to which $\mathbf{M}(\mathbb{X})\mathbb{X}$ belongs is invariant under transformation of $\mathbb{X}$ by $g \in \mathscr{G}$. We call the matrix $\mathbf{M}(\cdot)$ an *IWECS transformation*, and resulting the transformed data $\mathbf{M}(\mathbb{X})\mathbb{X}$ is the desired *IWECS*.

Whereas $\mathscr{G}$ concerns transformations on initial data sets $\mathbb{X}$ in $\mathbb{R}^d$ and represents a criterion for invariance, the equivalence relation $\mathscr{E}$ concerns the transformed data sets $\mathbf{M}(\mathbb{X})\mathbb{X}$ and represents a criterion for equivalent geometric structure. There are many possibilities for $(\mathscr{G}, \mathscr{E})$. In the example of Sect. 24.1, $\mathscr{G}$ consists of affine transformations and $\mathscr{E}$ represents data sets as having equivalent geometric structure if they differ only with respect to homogeneous scale change, coordinatewise sign changes, and translation.

For given $(\mathscr{G}, \mathscr{E})$, the challenge is to find a suitable $\mathbf{M}(\cdot)$ satisfying Definition 24.1. In the sequel, we consider the special case that $\mathscr{E}$ corresponds to invariance under a group of transformations $\mathscr{F}$ and denote the above framework by $(\mathscr{G}, \mathbf{M}(\cdot), \mathscr{F})$. Then a key criterion for finding a solution is provided by the following result, which follows immediately from Definition 24.1.

Theorem 24.1 *For given $(\mathscr{G}, \mathscr{F})$, a suitable IWECS transformation is given by any $\mathbf{M}(\cdot)$ such that, for any $g \in \mathscr{G}$, there exists $f_0 = f_0(g, \mathbb{X}) \in \mathscr{F}$ for which*

$$\mathbf{M}(g\mathbb{X})g\mathbb{X} = f_0\, \mathbf{M}(\mathbb{X})\mathbb{X}. \tag{24.1}$$

The following result provides a useful sufficient condition for (24.1) in the form of a structural requirement on the matrix $\mathbf{M}(\cdot)$ that in practice serves essentially as the definition of an IWECS transformation. The proof is immediate.

Corollary 24.1 *For given $(\mathscr{G}, \mathscr{F})$, a suitable IWECS transformation is given by any $\mathbf{M}(\cdot)$ such that, for any $g \in \mathscr{G}$, there exists $f_0 = f_0(g, \mathbb{X}) \in \mathscr{F}$ for which*

$$\mathbf{M}(g\mathbb{X}) = f_0\, \mathbf{M}(\mathbb{X})g^{-1}. \tag{24.2}$$

With $\mathscr{A}$ the set of all nonsingular $d \times d$ matrices, important choices for $\mathscr{G}$ are

$\mathscr{G}_0 = \{g : g\mathbb{X} = \mathbf{A}\mathbb{X},\ \mathbf{A} \in \mathscr{A}\}$ (nonsingular linear transformation),
$\mathscr{G}_1 = \{g : g\mathbb{X} = \mathbf{A}\mathbb{X} + \mathbf{b},\ \mathbf{A} \in \mathscr{A},\ \mathbf{b} \in \mathbb{R}^d\}$ (affine transformation).

For $\mathscr{F}$, key choices are

$\mathscr{D}_0 = \{f : f\mathbb{Y} = c\mathbb{Y},\ c > 0\}$ (homogeneous rescaling),
$\mathscr{D} = \{f : f\mathbb{Y} = \mathrm{diag}(c_1, \ldots, c_d)\mathbb{Y},\ c_i > 0,\ i = 1, \ldots, d\}$ (heterogeneous rescaling),
$\mathscr{J} = \{f : f\mathbb{Y} = \mathrm{diag}(c_1, \ldots, c_d)\mathbb{Y},\ c_i = \pm 1,\ i = 1, \ldots, d\}$ (heterogeneous sign changing),
$\mathscr{P} = \{f : f\mathbb{Y} = \mathbf{P}\mathbb{Y},\ \mathbf{P}$ is a permutation matrix$\}$ (permutation),
$\mathscr{U} = \{f : f\mathbb{Y} = \mathbf{U}\mathbb{Y},\ \mathbf{U}$ is an orthogonal matrix$\}$ (rotation and/or reflection).

The above $\mathscr{G}$ and $\mathscr{F}$ arise quite naturally in nonparametric multivariate inference, as discussed in Ilmonen et al. (2012), where $\mathscr{G}_1$ is especially emphasized. Their Eq. (24.4) corresponds to our Eq. (24.2) specialized to $\mathscr{G}_1$ and in that form is given as

their definition of what here we call an IWECS transformation. Let us also note that certain combinations of the above choices of $\mathscr{F}$ are of special interest, for example:

$\mathscr{F}_0 = \{f : f\mathbb{Y} = c\,\mathbb{Y} + \mathbf{b},\ c > 0,\ \mathbf{b} \in \mathbb{R}^d\}$
(translation, homogeneous rescaling),
$\mathscr{F}_1 = \{f : f\mathbb{Y} = c\,\mathrm{diag}(c_1, \ldots, c_d)\,\mathbb{Y} + \mathbf{b},\ c > 0,\ c_i = \pm 1,\ i = 1, \ldots, d,\ \mathbf{b} \in \mathbb{R}^d\}$
(translation, homogeneous rescaling, heterogeneous sign changing),
$\mathscr{F}_2 = \{f : f\mathbb{Y} = \mathrm{diag}(\pm c_1, \ldots, \pm c_d)\,\mathbb{Y} + \mathbf{b},\ c_i > 0,\ i = 1, \ldots, d,\ \mathbf{b} \in \mathbb{R}^d\}$
(translation, heterogeneous rescaling, heterogeneous sign changing),
$\mathscr{F}_3 = \{f : f\mathbb{Y} = \mathbf{U}\,(c\,\mathbb{Y} + \mathbf{b}),\ c > 0,\ \mathbf{U}$ orthogonal, $\mathbf{b} \in \mathbb{R}^d\}$
(translation, homogeneous rescaling, rotation, reflection),
$\mathscr{F}_4 = \{f : f\mathbb{Y} = \mathbf{U}\,(\mathrm{diag}(c_1, \ldots, c_d)\,\mathbb{Y} + \mathbf{b}),\ c_i > 0,\ i = 1, \ldots, d,\ \mathbf{U}$ orthogonal, $\mathbf{b} \in \mathbb{R}^d\}$
(translation, heterogeneous rescaling, rotation, reflection).

In particular, the example discussed in Sect. 24.1 concerns $\mathscr{G}_1$ and $\mathscr{F}_1$.

The property that a transformation is IWECS with respect to $(\mathscr{G}, \mathscr{F})$ becomes weaker if $\mathscr{F}$ acquires additional transformations. In this respect, let us note that $\mathscr{F}_0 \subset \mathscr{F}_1 \subset \mathscr{F}_2$ and $\mathscr{F}_0 \subset \mathscr{F}_3 \subset \mathscr{F}_4$ so that here the strongest case corresponds to $\mathscr{F} = \mathscr{F}_0$. Of course, still stronger is the ideal case that the equivalence relation $\mathscr{E}$ (i.e., the group $\mathscr{F}$) may be omitted and the invariance relative to $\mathscr{G}$ is strict: the transformed sets $\mathbf{M}(g\mathbb{X})g\mathbb{X}$, $g \in \mathscr{G}$, are identical without qualification and $\mathbf{M}(\cdot)$ is a purely ICS transformation. Generally, however, this aspiration is too stringent and must be relaxed, adopting an equivalence criterion that is as narrow as possible.

24.3 TR Matrices as IWECS Transformations

In seeking IWECS transformations relative to the popular affine group $\mathscr{G}_1$, one may inquire whether widely used standardizing transformations such as the inverse square roots of scatter matrices suffice for this purpose. That is, more precisely, may a transformation-retransformation (TR) matrix serve as an IWECS matrix? As shown in Serfling (2010), the answer is negative except for some very special cases that exclude popular ones. Hence it becomes of interest to explore what "minimal" $\mathscr{F}$ suffices for an arbitrary TR transformation to serve as an IWECS transformation relative to $(\mathscr{G}_1, \mathscr{F})$. In Sect. 24.3.1 we review the definition of TR matrices, and in Sects. 24.3.2 and 24.3.3 we develop explicit answers to this question.

24.3.1 Definition of TR Matrices

A *transformation-retransformation (TR)* matrix is a positive definite $d \times d$ matrix $\mathbf{M}(\mathbb{X})$ (not necessarily symmetric) such that, for $\mathbb{Y} = \mathbf{A}\mathbb{X} + \mathbf{b}$ with any nonsingular $\mathbf{A}$ and any $\mathbf{b}$,

$$\mathbf{A}^\top \mathbf{M}(\mathbb{Y})^\top \mathbf{M}(\mathbb{Y})\mathbf{A} = k_2 \mathbf{M}(\mathbb{X})^\top \mathbf{M}(\mathbb{X}),$$

with $k_2 = k_2(\mathbf{A}, \mathbf{b}, \mathbb{X})$ a positive scalar function of $\mathbf{A}$, $\mathbf{b}$, and $\mathbb{X}$. Such TR matrices are equivalently given by factorizations of weak covariance (WC) matrices, i.e., via

$$\mathbf{C}(\mathbb{X}) = (\mathbf{M}(\mathbb{X})^\top \mathbf{M}(\mathbb{X}))^{-1},$$

where the symmetric positive definite $d \times d$ WC matrix $\mathbf{C}(\mathbb{X})$ satisfies

$$\mathbf{C}(\mathbb{Y}) = k_1 \mathbf{A}\mathbf{C}(\mathbb{X})\mathbf{A}^\top,$$

with $k_1 = k_1(\mathbf{A}, \mathbf{b}, \mathbb{X})$ a positive scalar function of $\mathbf{A}$, $\mathbf{b}$, and $\mathbb{X}$. For $k_1 = 1$, $\mathbf{C}(\mathbb{X})$ is a strict "covariance" matrix. Typical standardizations of data $\mathbb{X}$ for various purposes are given by $\mathbf{M}(\mathbb{X})\mathbb{X}$. See Serfling (2010) and Ilmonen et al. (2012) for detailed discussion and examples.

24.3.2 TR Matrices as IWECS Transformations

We now explore whether such an $\mathbf{M}(\cdot)$ can be IWECS. In particular, Serfling (2010) shows, in different notation, that any IWECS transformation relative to $(\mathscr{G}_1, \mathscr{F}_1)$ is TR, but not conversely, one counterexample being the popular (Tyler 1987) TR matrix. But is there a broader $\mathscr{F}$ for which *any TR matrix* is in fact IWECS? The following result answers this in the affirmative, for both $\mathscr{G}_1$ and $\mathscr{G}_0$, with $\mathscr{F} = \mathscr{F}_3$.

Theorem 24.2 *Every TR matrix is IWECS relative to* $(\mathscr{G}_1, \mathscr{F}_3)$ *and also to* $(\mathscr{G}_0, \mathscr{F}_3)$.

Proof

(i) Let us first consider $(\mathscr{G}_1, \mathscr{F}_3)$. Let $g \in \mathscr{G}_1$ be given by $g\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{b}$ for some nonsingular $\mathbf{A}$ and any $\mathbf{b}$. It is shown in Serfling (2010), Lemma 5.1, that, for any TR matrix $\mathbf{M}(\cdot)$, and for $\mathbb{Y} = \mathbf{A}\mathbb{X} + \mathbf{b}$ with any nonsingular $\mathbf{A}$ and any $\mathbf{b}$, the matrix

$$\mathbf{U}_0 = \mathbf{U}_0(\mathbf{A}, \mathbf{b}, \mathbb{X}) = k_2(\mathbf{A}, \mathbf{b}, \mathbb{X})^{1/2} (\mathbf{M}(\mathbb{Y})^\top)^{-1} (\mathbf{A}^\top)^{-1} \mathbf{M}(\mathbb{X})^\top$$

is orthogonal. Then we readily obtain

$$\mathbf{M}(\mathbb{Y}) = k_2^{1/2} \mathbf{U}_0 \mathbf{M}(\mathbb{X}) \mathbf{A}^{-1} \tag{24.3}$$

and in turn

$$\mathbf{M}(\mathbb{Y})\mathbb{Y} = k_2^{1/2} \mathbf{U}_0 [\mathbf{M}(\mathbb{X})\mathbb{X} + \mathbf{M}(\mathbb{X})\mathbf{A}^{-1}\mathbf{b}] = f_0 \mathbf{M}(\mathbb{X})\mathbb{X}, \tag{24.4}$$

where $f_0 = f_0(\mathbf{A}, \mathbf{b}, \mathbb{X})$ represents translation of $\mathbf{M}(\mathbb{X})\mathbb{X}$ by the constant $\mathbf{M}(\mathbb{X})\mathbf{A}^{-1}\mathbf{b}$, followed by homogeneous scale change by $k_2(\mathbf{A}, \mathbf{b}, \mathbb{X})$ and then

rotation/reflection by orthogonal $\mathbf{U}_0$. Thus $f_0 \in \mathscr{F}_3$ and Eq. (24.1) in Theorem 24.1 is satisfied for the given $g \in \mathscr{G}_1$.

(ii) For $(\mathscr{G}_0, \mathscr{F}_3)$, the proof is similar. □

Theorem 24.2 shows explicitly the precise strengths and limitations of TR matrices as IWECS transformations. We can apply this result through various straightforward corollaries, for example the following.

Corollary 24.2 *If an $\mathbb{R}^m$-valued statistic $\mathbf{Q}(\mathbb{X})$ is invariant with respect to $\mathscr{F}_3$, then its evaluation at a TR-based IWECS $\mathbf{M}(\mathbb{X})\mathbb{X}$ relative to either $(\mathscr{G}_1, \mathscr{F}_3)$ or $(\mathscr{G}_0, \mathscr{F}_3)$ is invariant with respect to $\mathscr{G}_1$ or $\mathscr{G}_0$, respectively.*

Example 24.1 Invariance of spatial outlyingness function. The *spatial outlyingness function* (Serfling 2010) is defined as

$$O_S(\mathbf{x}, \mathbb{X}) = \|\mathbf{R}_S(\mathbf{x}, \mathbb{X})\|, \ \mathbf{x} \in \mathbb{R}^d,$$

where $\mathbf{R}_S(\mathbf{x}, \mathbb{X})$ is the spatial centered rank function (Oja 2010) in $\mathbb{R}^d$ given by

$$\mathbf{R}_S(\mathbf{x}, \mathbb{X}) = n^{-1} \sum_{i=1}^{n} \mathbf{S}(\mathbf{x} - \mathbf{X}_i), \ \mathbf{x} \in \mathbb{R}^d,$$

with $\mathbf{S}(\mathbf{y})$ the *spatial sign function* (or *unit vector function*) in $\mathbb{R}^d$ given by

$$\mathbf{S}(\mathbf{y}) = \begin{cases} \frac{\mathbf{y}}{\|\mathbf{y}\|}, \ \mathbf{y} \in \mathbb{R}^d, \mathbf{y} \neq \mathbf{0}, \\ \mathbf{0}, \quad \mathbf{y} = \mathbf{0}. \end{cases}$$

It is readily checked that $\mathbf{R}_S(\mathbf{x}, \mathbb{X})$ is *translation and homogeneous scale invariant* and *orthogonally equivariant*:

$$\mathbf{R}_S(\mathbf{x} - \mathbf{b}, \mathbb{X} - \mathbf{b}) = \mathbf{R}_S(\mathbf{x}, \mathbb{X}),$$

$$\mathbf{R}_S(c\mathbf{x}, c\mathbb{X}) = \mathbf{R}_S(\mathbf{x}, \mathbb{X}),$$

$$\mathbf{R}_S(\mathbf{U}\mathbf{x}, \mathbf{U}\mathbb{X}) = \mathbf{U}\mathbf{R}_S(\mathbf{x}, \mathbb{X}).$$

Then $O_S(\mathbf{x}, \mathbb{X})$ is *translation, homogeneous scale, and orthogonally invariant, i.e., invariant with respect to the group $\mathscr{F}_3$*, but is not affine invariant. However, fully affine invariant versions are immediately obtained via Corollary 24.2: For any TR matrix $\mathbf{M}(\cdot)$, the so-called *TR spatial outlyingness function* corresponding to $\mathbf{M}(\cdot)$,

$$O_S^{(\mathrm{TR})}(\mathbf{x}, \mathbb{X}) = O_S(\mathbf{M}(\mathbb{X})\mathbf{x}, \mathbf{M}(\mathbb{X})\mathbb{X}), \ \mathbf{x} \in \mathbb{R}^d,$$

is *affine invariant*. □

24.3.3 Special Types of TR Matrix for IWECS Transformations

It would be desirable to have the result of Theorem 24.2 for a smaller choice of $\mathscr{F}$ than $\mathscr{F}_3$. In this vein, we ask: *What additional property is required of a TR matrix* $\mathbf{M}(\cdot)$ *in order for it to serve as an IWECS transformation relative to* $\mathscr{G}_1$ *for a choice of* $\mathscr{F}$ *narrower than* $\mathscr{F}_3$*?*

A clue is given by Eq. (24.3) in the proof of Theorem 24.2. If we simply require that $\mathbf{M}(\cdot)$ satisfy this equation *without* the factor $\mathbf{U}_0$, then Eq. (24.4) would hold without the presence of $\mathbf{U}_0$, yielding the following very strong conclusion.

Theorem 24.3 *Let* $\mathbf{M}(\cdot)$ *be a TR matrix such that, for* $\mathbb{Y} = \mathbf{A}\mathbb{X} + \mathbf{b}$ *as above,*

$$\mathbf{M}(\mathbb{Y}) = k_2^{1/2}\mathbf{M}(\mathbb{X})\mathbf{A}^{-1}, \tag{24.5}$$

with $k_2 = k_2(\mathbf{A}, \mathbf{b}, \mathbb{X})$ *as in the definition of the given TR matrix. Then* $\mathbf{M}(\cdot)$ *is IWECS relative to* $(\mathscr{G}_1, \mathscr{F}_0)$ *and to* $(\mathscr{G}_0, \mathscr{F}_0)$.

An analogue of Corollary 24.2 is

Corollary 24.3 *If an* $\mathbb{R}^m$*-valued statistic* $\mathbf{Q}(\mathbb{X})$ *is invariant with respect to* $\mathscr{F}_0$*, then its evaluation at a TR-based IWECS* $\mathbf{M}(\mathbb{X})\mathbb{X}$ *for* $\mathbf{M}(\cdot)$ *satisfying (24.5) is, relative to either* $(\mathscr{G}_1, \mathscr{F}_0)$ *or* $(\mathscr{G}_0, \mathscr{F}_0)$*, invariant with respect to* $\mathscr{G}_1$ *or* $\mathscr{G}_0$*, respectively.*

In comparison with Corollary 24.2, Corollary 24.3 requires more of the TR matrix but yields a stronger conclusion by allowing $\mathscr{F}_0$ instead of $\mathscr{F}_3$.

Example 24.2 Invariance of projection outlyingness with finitely many projections. With ν the median and η the MAD (median absolute deviation from the median), the well-known *projection outlyingness function* given by

$$O_P(\mathbf{x}, \mathbb{X}) = \sup_{\|\mathbf{u}\| = 1} \left| \frac{\mathbf{u}^\top \mathbf{x} - \nu(\mathbf{u}^\top \mathbb{X})}{\eta(\mathbf{u}^\top \mathbb{X})} \right|, \ \mathbf{x} \in \mathbb{R}^d, \tag{24.6}$$

represents the worst case scaled deviation outlyingness of projections of $\mathbf{x}$ onto lines. It is affine invariant, highly masking robust (Dang and Serfling 2010), and does not impose ellipsoidal contours as does the very popular Mahalanobis distance outlyingness function, which also is affine invariant. However, $O_P(\mathbf{x}, \mathbb{X})$ is highly computational, and to overcome this burden Serfling and Mazumder (2013) develop and study a modified version entailing only finitely many selected projections, say $\Delta = \{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$, i.e.,

$$O_P^{(\Delta)}(\mathbf{x}, \mathbb{X}) = \sup_{\mathbf{u} \in \Delta} \left| \frac{\mathbf{u}^\top \mathbf{x} - \nu(\mathbf{u}^\top \mathbb{X})}{\eta(\mathbf{u}^\top \mathbb{X})} \right|, \ \mathbf{x} \in \mathbb{R}^d. \tag{24.7}$$

However, $O_P^{(\Delta)}(\mathbf{x}, \mathbb{X})$ with finite Δ is no longer affine invariant. Nor is it orthogonally invariant, so Corollary 24.2 is inapplicable and thus an arbitrary TR version does not

achieve affine invariance. On the other hand, simply using invariance of $O_P^{(\Delta)}(\mathbf{x}, \mathbb{X})$ with respect to $\mathscr{F}_0$, it follows by Corollary 24.3 that any TR version

$$O_P^{(\Delta,TR)}(\mathbf{x}, \mathbb{X}) = O_P^{(\Delta)}(\mathbf{M}(\mathbb{X})\mathbf{x}, \mathbf{M}(\mathbb{X})\mathbb{X}), \ \mathbf{x} \in \mathbb{R}^d,$$

with $\mathbf{M}(\cdot)$ satisfying (24.5) is indeed affine invariant. Of course, standardizing by $\mathbf{M}(\cdot)$ introduces a further computational issue, and Serfling and Mazumder (2013) also develop computationally attractive choices of $\mathbf{M}(\cdot)$ satisfying (24.5). □

Remark 24.1 A TR matrix satisfying the special condition (24.5) is distinguished as a "strong invariant coordinate system" (SICS) transformation in Serfling (2010) and Ilmonen et al. (2012), where also other results like Theorem 24.3 are seen corresponding to replacement of $\mathbf{U}_0$ in (24.3) by possibilities other than simply the identity matrix and hence corresponding to $\mathscr{F}$ larger than $\mathscr{F}_0$. □

Remark 24.2 Let us compare condition (24.5) with the somewhat similar condition given by (24.2), which in the present setting would be expressed as

$$\mathbf{M}(\mathbb{Y}) = f_0 \mathbf{M}(\mathbb{X}) g^{-1} \tag{24.8}$$

for any given $g \in \mathscr{G}$ for some related $f_0 \in \mathscr{F}_0$. For the case $\mathscr{G} = \mathscr{G}_1$, let g be given by $g\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{b}$ for some nonsingular $\mathbf{A}$ and any $\mathbf{b}$. Now it is readily checked that the transformation g^{-1} consists of translation by $-\mathbf{b}$ followed by application of $\mathbf{A}^{-1}$, or equivalently application of $\mathbf{A}^{-1}$ followed by translation by $-\mathbf{A}^{-1}\mathbf{b}$. *Thus (24.5) is equivalent to (24.2) with suitable choice of* $f_0 \in \mathscr{F}_0$. The argument for $\mathscr{G} = \mathscr{G}_0$ is similar. In dealing with $\mathscr{G} = \mathscr{G}_1$, the use of (24.5) is more direct and convenient. □

24.4 Some Connections with Maximal Invariants

A natural "invariance principle" is that artifacts of the data $\mathbb{X}$ which are invariant without qualification relative to a group $\mathscr{G}$ of transformations should be functions of a suitable "maximal invariant" statistic that constitutes a labeling of the orbits of $\mathscr{G}$. See Lehmann and Romano (2005, Sect. 6.2), for elaboration. For a data set $\mathbb{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ of observations in $\mathbb{R}^d$, a maximal invariant is obtained via some suitable matrix-valued transformation $\mathbf{B}(\mathbb{X})$ applied to $\mathbb{X}$, producing $\mathbf{B}(\mathbb{X})\mathbb{X}$ as the desired maximal invariant.

In this case, an ICS $\mathbf{M}(\mathbb{X})\mathbb{X}$ relative to $\mathscr{G}$ should be expressible as a function of $\mathbf{B}(\mathbb{X})\mathbb{X}$. However, this need not be true for an IWECS, of course. In Sect. 24.4.1, relative to $\mathscr{G}_0$ and $\mathscr{G}_1$, we exhibit connections between ICS transformations and the pertinent maximal invariant transformations. These connections are exploited in Sect. 24.4.2 to "discover" from some classical results a new approach toward construction of ICS and IWECS transformations. In Sect. 24.5 the connections are extended to the case of an arbitrary $\mathscr{G}$.

24.4.1 Connections in the Case of Groups $\mathscr{G}_0$ and $\mathscr{G}_1$

With reference to the groups $\mathscr{G}_0$ and $\mathscr{G}_1$, the following results of Ilmonen et al. (2012, Theorem 3.1), connect maximal invariants with TR matrices $\mathbf{M}(\cdot)$ satisfying a strong special case of Eq. (24.5), namely

$$\mathbf{M}(\mathbb{Y}) = \mathbf{M}(\mathbb{X})\mathbf{A}^{-1}, \tag{24.9}$$

where $\mathbb{Y}$ denotes $\mathbf{A}\mathbb{X}$ in the case of $\mathscr{G}_0$ and $\mathbf{A}\mathbb{X} + \mathbf{b}$ in the case of $\mathscr{G}_1$. They show under (24.9) that

(i) $\mathbf{M}(\mathbb{X})\mathbb{X}$ is a maximal invariant under $\mathscr{G}_0$,
(ii) $\mathbf{M}(\mathbb{X})(\mathbb{X} - T(\mathbb{X}))$ is a maximal invariant under $\mathscr{G}_1$, for any location statistic $T(\mathbb{X})$.

Of course, in view of (24.9), $\mathbf{M}(\cdot)$ is an IWECS transformation. Also, note that under (24.9) we have $\mathbf{M}(\mathbb{X}) = \mathbf{M}(\mathbb{X} - T(\mathbb{X}))$, and thus the maximal invariant in (ii) may also be written as $\mathbf{M}(\mathbb{X} - T(\mathbb{X}))(\mathbb{X} - T(\mathbb{X}))$.

In case (i), $\mathbf{M}(\cdot)$ is a very strong special case of IWECS transformation, namely a pure ICS transformation without qualification by an equivalence relation, for we have $\mathbf{M}(g\mathbb{X})g\mathbb{X} = \mathbf{M}(\mathbb{X})\mathbb{X}$, $g \in \mathscr{G}_0$. Note that under merely (24.5) instead of the strengthening to (24.9), we have by Theorem 24.3 that $\mathbf{M}(\cdot)$ is an IWECS transformation relative to $(\mathscr{G}_1, \mathscr{F}_0)$, a slightly weaker conclusion although still quite strong, and the IWECS $\mathbf{M}(\mathbb{X})\mathbb{X}$ is no longer a maximal invariant.

In case (ii), and even under merely (24.5), we have that $\mathbf{M}(\cdot)$ is IWECS relative to $(\mathscr{G}_1, \mathscr{F}_0)$, as per Theorem 24.3. However, the IWECS $\mathbf{M}(\mathbb{X})\mathbb{X}$ is not a maximal invariant. Consequently, under (24.9), $\mathbf{M}(\cdot)$ is closely associated with both obtaining an IWECS and obtaining a maximal invariant, although neither solution directly yields the other. Since typical TR matrices do not satisfy (24.9), special types are required.

Particular constructions of $\mathbf{M}(\cdot)$ satisfying (24.9) with reference to $\mathscr{G}_0$ and $\mathscr{G}_1$ have been developed and applied to obtain affine invariant multivariate sign and angle tests and affine equivariant multivariate coordinatewise and spatial medians, in a series of papers by Chaudhuri and Sengupta (1993), Chakraborty and Chaudhuri (1996), and Chakraborty et al. (1998). Further approaches are treated in Serfling (2010), and Ilmonen et al. (2010, 2012), covering a range of applications and exploring the formal properties of these transformations. Treatments are carried out in the setting of complex valued independent component analysis by Ilmonen (2013) and in the setting of supervised invariant coordinate selection by Liski et al. (2014).

24.4.2 Some Pertinent Classical Results

Maximal invariant statistics relative to $\mathscr{G}_0$ and $\mathscr{G}_1$ have been treated in detail as early as Lehmann (1959), and those results are pertinent here. In particular, for $\mathbb{X}$ a $d \times n$ matrix of n column d-vectors, and relative to the group $\mathscr{G}_0$, Lehmann (1959) derives

the maximal invariant

$$\mathbb{P} = \mathbb{X}^\top (\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X},$$

which corresponds to $\mathbf{M}_0(\mathbb{X})\mathbb{X}$ with

$$\mathbf{M}_0(\mathbb{X}) = \mathbb{X}^\top (\mathbb{X}\mathbb{X}^\top)^{-1}.$$

We readily find that $\mathbf{M}_0(\cdot)$ satisfies (24.9) and hence is both an ICS and a maximal invariant transformation.

Three noteworthy aspects of the Lehmann maximal invariant $\mathbb{P}$ are as follows:

1. $\mathbf{M}_0(\mathbb{X})$ is more directly computed than existing ICS matrices relative to $\mathscr{G}_1$.
2. $\mathbb{P}$ has interesting geometric interpretations as discussed in Lehmann (1959).
3. $\mathbb{P}$ is $n \times n$ rather than $d \times n$ as would be $\mathbf{M}_0(\mathbb{X})\mathbb{X}$ were $\mathbf{M}_0(\cdot)$ a $d \times d$ TR matrix satisfying (24.9). However, as easily seen, assuming the rows of $\mathbf{M}_0(\mathbb{X})$ are linearly independent as should hold with probability 1, any d rows of $\mathbf{M}_0(\mathbb{X})$ form a $d \times d$ TR matrix $\mathbf{M}_1(\cdot)$, say, also satisfying (24.9) and thus yield what we might call a *minimal dimension* maximal invariant $\mathbb{P}_0 = \mathbf{M}_1(\mathbb{X})\mathbb{X}$, say.

Remark 24.3 Note that $\mathbb{P}_0$ (a) serves as an ICS relative to $\mathscr{G}_0$, (b) serves as a maximal invariant relative to $\mathscr{G}_1$ via

$$\mathbf{M}_1(\mathbb{X})(\mathbb{X} - T(\mathbb{X})) = \mathbf{M}_1(\mathbb{X} - T(\mathbb{X}))(\mathbb{X} - T(\mathbb{X})),$$

and (c) serves as an IWECS relative to $(\mathscr{G}_1, \mathscr{F}_0)$. The reduction of the "full" maximal invariant $\mathbb{P}$ to the minimal dimension version $\mathbb{P}_0$ gives up some data, but only what is redundant of that which is retained, as far as a labeling of orbits is concerned. It should be noted that the computational burden posed by $\mathbb{P}$ and $\mathbb{P}_0$ is relatively light. Full investigation of $\mathbb{P}$ and $\mathbb{P}_0$ is deferred to a future study. □

24.5 Extensions for General $\mathbb{X}$ and General $\mathscr{G}$

In the present paper we have focused on $\mathscr{X} = \mathbb{R}^d$ and $\mathscr{G} = \mathscr{G}_0$ and $\mathscr{G}_1$. However, Definition 24.1 can immediately be formulated more generally, allowing the data $\mathbb{X}$ to be observations from any space $\mathscr{X}$ and taking $\mathbf{M}$ to be a data based operator on elements of $\mathscr{X}$.

Also, the connections (i) and (ii) of Sect. 24.4.1 regarding maximal invariance under (24.9) have a completely general extension, corresponding to a tightening of (24.2) in the same way that (24.9) tightens (24.5), as follows.

Theorem 24.4 *For any group $\mathscr{G}$ of transformations on data sets $\mathbb{X}$ from any space $\mathscr{X}$, let $\mathbf{M}(\cdot)$ be such that $\mathbf{M}(\mathbb{X})$ itself belongs to $\mathscr{G}$ for any data set $\mathbb{X}$ and suppose that*

$$\mathbf{M}(g\mathbb{X}) = \mathbf{M}(\mathbb{X})g^{-1}. \tag{24.10}$$

Then $\mathbf{M}(\mathbb{X})\mathbb{X}$ is both an ICS and a maximal invariant with respect to $\mathscr{G}$.

Proof Invariance of $\mathbf{M}(\mathbb{X})\mathbb{X}$ follows immediately from (24.10). Now suppose that $\mathbf{M}(\mathbb{X})\mathbb{X} = \mathbf{M}(\mathbb{X}^*)\mathbb{X}^*$ for two data sets $\mathbb{X}$ and $\mathbb{X}^*$. Then $\mathbb{X}^* = [\mathbf{M}(\mathbb{X}^*)^{-1}\mathbf{M}(\mathbb{X})]\mathbb{X} = g^*\mathbb{X}$, where $g^* = \mathbf{M}(\mathbb{X}^*)^{-1}\mathbf{M}(\mathbb{X}) \in \mathscr{G}$. Hence $\mathbb{X}$ and $\mathbb{X}^*$ lie in the same orbit of $\mathscr{G}$, establishing maximality. □

Statistical inference procedures which are invariant or equivariant with respect to some group $\mathscr{G}$ can be obtained by evaluating suitable preliminary versions at some appropriate functions either of an IWECS or of a maximal invariant, whichever is more convenient. In light of Theorem 24.4, these constructions and studies may be explored in greater generality than for $\mathscr{X} = \mathbb{R}^d$. For example, a potential application of the IWECS framework arises in the study of similarity between time series with invariance to (various combinations of) the distortions of warping, uniform scaling, offset, amplitude scaling, phase, occlusions, uncertainty, and wandering baseline. This and other applications are being pursued in separate investigations.

Acknowledgements The author thanks G.L. Thompson for encouragement and two anonymous reviewers for comments and suggestions helpful to improving the paper. Also, support by NSF Grant DMS-1106691 is greatly appreciated.

References

Chakraborty, B., Chaudhuri, P.: On a transformation and re-transformation technique for constructing an affine equivariant multivariate median. Proc. Am. Math. Soc. **124**, 2539–2547 (1996)

Chakraborty, B., Chaudhuri, P., Oja, H.: Operating transformation retransformation on spatial median and angle test. Stat. Sin. **8**, 767–784 (1998)

Chaudhuri, P., Sengupta, D.: Sign tests in multidimension: inference based on the geometry of the data cloud. J. Am. Stat. Assoc. **88**, 1363–1370 (1993)

Dang, X., Serfling, R.: Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties. J. Stat. Plan. Infer. **140**, 198–213 (2010)

Ilmonen, P.: On asymptotic properties of the scatter matrix based estimates for complex valued independent component analysis. Stat. Probab. Lett. **83**, 1219–1226 (2013)

Ilmonen, P., Nevalainen, J., Oja, H.: Characteristics of multivariate distributions and the invariant coordinate system. Stat. Probab. Lett. **80**, 1844–1853 (2010)

Ilmonen, P., Oja, H., Serfling, R.: On invariant coordinate system (ICS) functionals. Int. Stat. Rev. **80**, 93–110 (2012)

Lehmann, E.L.: Testing Statistical Hypotheses. Wiley, New York (1959)

Lehmann, E.L., Romano, J.P.: Testing Statistical Hypotheses, 3rd edn. Springer, New York (2005)

Liski, E., Nordhausen, K., Oja, H.: On supervised coordinate selection. Stat. J. Theor. Appl. Stat. **48**, 711–731 (2014)

Nordhausen, K.: On invariant coordinate selection and nonparametric analysis of multivariate data. Ph.D. Dissertation, University of Tampere (2008)
Oja, H.: Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks. Springer, New York (2010)
Serfling, R.: Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardization. J. Nonparametr. Stat. **22**, 915–936 (2010)
Serfling, R., Mazumder, S.: Computationally easy outlier detection via projection pursuit with finitely many directions. J. Nonparametr. Stat. **25**, 447–461 (2013)
Tyler, D.E.: A distribution-free M-estimator of scatter. Ann. Stat. **15**, 234–251 (1987)
Tyler, D.E., Critchley, F., Dümbgen, L., Oja, H.: . Invariant co-ordinate selection. J. R. Stat. Soc. Ser. B **71**, 549–592 (2009)

Chapter 25
Alternative Diagonality Criteria for SOBI

Jari Miettinen

Abstract Blind source separation (BSS) is a multivariate data analysis method, whose roots are in the signal processing community. BSS is applied in diverse fields, including, for example, brain imaging and economic time series analysis. In the BSS model there are interesting latent uncorrelated variables, and the aim is to estimate the latent variables from multiple linear combinations of them. In this article we assume that these variables are weakly stationary time series, and we consider estimation methods which are based on approximate joint diagonalization of autocovariance matrices. In the popular SOBI estimator, a set of matrices is most diagonal when the sum of squares of their diagonal elements is maximal. Here we investigate other criteria to measure the diagonality of matrices. Applying both asymptotic results and simulations, we will study how the use of different diagonality measures affects the separation performance. Also, a method to choose the measure optimally based on data is proposed.

Keywords Approximate joint diagonalization • Blind source separation • Time series

25.1 Introduction

The blind source separation (BSS) model in its most simple form is written as $\mathbf{x} = \boldsymbol{\Omega}\mathbf{z} + \boldsymbol{\mu}$, where $\mathbf{x}$ is the observed p-variate vector, $\boldsymbol{\Omega}$ is a full rank $p \times p$ mixing matrix, and $\mathbf{z}$ is a p-variate latent source vector. The p-variate location vector $\boldsymbol{\mu}$ is most often considered as a nuisance parameter, when the aim is to estimate the mixing matrix $\boldsymbol{\Omega}$, or the unmixing matrix $\boldsymbol{\Gamma} = \boldsymbol{\Omega}^{-1}$, based on the observations $\mathbf{x}_1, \ldots, \mathbf{x}_T$.

The identifiability of the unmixing matrix depends on the distribution of the source vector $\mathbf{z}$. In independent component analysis (ICA) the components of $\mathbf{z}$ are mutually independent, and the unmixing matrix is identifiable if at most one of the components is Gaussian (Comon 1994). In ICA the possible temporal dependence

J. Miettinen (✉)
Department of Mathematics and Statistics, University of Jyväskylä, 40014 Jyväskylä, Finland
e-mail: jari.p.miettinen@jyu.fi

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_25

is usually disregarded, but there are other branches of BSS, which purely employ the temporal structure of the observations. Then uncorrelatedness of the components is enough and multiple Gaussian components are allowed. Instead, the components must have mutually different temporal structures. BSS methods for time series still divide into two categories, one for weakly stationary and the other for nonstationary time series.

In this paper we consider the model

$$\mathbf{x}_t = \boldsymbol{\Omega}\mathbf{z}_t, \quad t = 0, \pm 1, \pm 2, \ldots \tag{25.1}$$

where the components of $\mathbf{z} = (\mathbf{z}_t)_{t=0,\pm 1,\pm 2}$ are uncorrelated weakly stationary time series which satisfy the following assumptions.

(A1) $E(\mathbf{z}_t) = \mathbf{0}$,
(A2) $E(\mathbf{z}_t\mathbf{z}_t') = \mathbf{I}_p$, and
(A3) $E(\mathbf{z}_t\mathbf{z}_{t+k}') = E(\mathbf{z}_{t+k}\mathbf{z}_t') = \boldsymbol{\Lambda}_k$ is diagonal for all $k = 1, 2, \ldots$, and for all pairs $i \neq j$ there is such $k \geq 1$ that $(\boldsymbol{\Lambda}_k)_{ii} \neq (\boldsymbol{\Lambda}_k)_{jj}$.

This is a semiparametric model since the distribution of $\mathbf{z}$ remains unspecified. The Assumption (A2) that $\mathbf{z}$ is standardized is motivated by the fact that otherwise the scales of the columns of $\boldsymbol{\Omega}$ and the scales of the components of $\mathbf{z}$ would be confounded in Model (25.1). The signs and order of the columns of $\boldsymbol{\Omega}$ cannot be estimated with these assumptions, but they are of less importance. An unmixing matrix functional $\boldsymbol{\Gamma}$ is said to be affine equivariant, if for $\mathbf{x}$ and $\mathbf{x}^* = \boldsymbol{A}\mathbf{x}$ for some non-singular $p \times p$ matrix $\mathbf{A}$, the estimates $\boldsymbol{\Gamma}_{\mathbf{x}}\mathbf{x}$ and $\boldsymbol{\Gamma}_{\mathbf{x}^*}\mathbf{x}^*$ are the same up to order and sign changes of the components. Affine equivariance is a desirable property which implies that the estimation accuracy does not depend on the mixing matrix $\boldsymbol{\Omega}$. Notice that we assume that the location vector $\boldsymbol{\mu} = \mathbf{0}$ is known. With our assumptions in the following sections, this will not be a restriction.

Probably the most popular functionals in Model (25.1) are based on the use of autocovariance matrices. Either on simultaneous diagonalization of two matrices (Tong et al. 1990; Ziehe and Müller 1998) or on approximate joint diagonalization of several matrices (Belouchrani et al. 1997). Indeed, more than two matrices can be diagonalized only approximately, and then joint diagonalization means maximization of some selected diagonality measure. Usually it is the sum of squares of the diagonal elements, but in this paper we consider also other choices for the diagonality measure.

The outline of the paper is the following. In Sect. 25.2 we will recall how joint diagonalization of autocovariance matrices have been used to solve the BSS problem, and define a family of new functionals. The asymptotical properties of the new functionals are presented in Sect. 25.3, first in the general case and then in more detail for linear processes. A new functional, which utilizes the asymptotical results, is proposed. The finite-sample properties and the asymptotic efficiency of the different estimators are compared in Sect. 25.4 with simulation studies. Finally, Sect. 25.5 provides a short discussion.

25.2 BSS and Autocovariance Matrices

25.2.1 Functionals Based on Simultaneous Diagonalization of Two Matrices

Write

$$\mathbf{S}_k = E(\mathbf{x}_t \mathbf{x}_{t+k}') = \boldsymbol{\Omega} \boldsymbol{\Lambda}_k \boldsymbol{\Omega}', \quad k = 0, 1, 2, \ldots$$

for the autocovariance matrices of $\mathbf{x}$ from the BSS model.

The first BSS estimator which utilizes autocovariance matrices was the so-called AMUSE (Algorithm for Multiple Unknown Signals Extraction) estimator proposed by Tong et al. (1990). The AMUSE unmixing matrix functional $\boldsymbol{\Gamma}_k$, for some lag k, satisfies

$$\boldsymbol{\Gamma}_k \mathbf{S}_0 \boldsymbol{\Gamma}_k' = \mathbf{I}_p \quad \text{and} \quad \boldsymbol{\Gamma}_k \mathbf{S}_k \boldsymbol{\Gamma}_k' = \boldsymbol{\Lambda}_k,$$

where $\boldsymbol{\Lambda}_k$ is a diagonal matrix with the diagonal elements in a decreasing order.

The functional $\boldsymbol{\Gamma}_k$ is affine equivariant, and it produces consistent estimates if the diagonal elements of $\boldsymbol{\Lambda}_k$ are distinct. AMUSE is computationally very simple method, but the choice of the lag is a problem (Miettinen et al. 2012). Ziehe and Müller (1998) suggested replacing the single autocovariance matrix $\mathbf{S}_k$ by the average $K^{-1} \sum_k S_k$ of K autocovariance matrices.

25.2.2 Functionals Based on Approximate Joint Diagonalization of Several Matrices

In terms of the estimation precision, it is better to conduct the joint diagonalization of several matrices as in the SOBI (Second Order Blind Identification) algorithm (Belouchrani et al. 1997).

The SOBI unmixing matrix functional $\boldsymbol{\Gamma}$ maximizes

$$\sum_{k=1}^{K} ||\text{diag}(\boldsymbol{\Gamma} \mathbf{S}_k \boldsymbol{\Gamma}')||^2 = \sum_{j=1}^{p} \sum_{k=1}^{K} (\boldsymbol{\gamma}_j' \mathbf{S}_k \boldsymbol{\gamma}_j)^2$$

under the constraint $\boldsymbol{\Gamma} \mathbf{S}_0 \boldsymbol{\Gamma}' = \mathbf{I}_p$. Here $\text{diag}(\mathbf{A})$ is a $p \times p$ diagonal matrix with the same diagonal elements as A and $\|\cdot\|$ is the matrix (Frobenius) norm. The constraint implies that the estimated source components are uncorrelated and standardized. It relates to the so-called whitening procedure, which is the first step in most of the BSS methods. In whitening the observed variable $\mathbf{x}$ is transformed to $\mathbf{y}_t = \mathbf{S}_0^{-1/2} \mathbf{x}_t$ (let $\mathbf{S}_0^{-1/2}$ be the symmetric matrix). Then $E(\mathbf{y}_t \mathbf{y}_t') = \mathbf{I}_p$ and $\mathbf{z}_t = \mathbf{U} \mathbf{y}_t$ for some

orthogonal matrix $\mathbf{U}$. The SOBI functional can be defined equivalently as $\boldsymbol{\Gamma} = \mathbf{U}\mathbf{S}_0^{-1/2}$, where $\mathbf{U}$ is the orthogonal matrix which maximizes

$$\sum_{k=1}^{K} ||\text{diag}(\mathbf{U}\mathbf{R}_k\mathbf{U}')||^2, \tag{25.2}$$

and $\mathbf{R}_k = \mathbf{S}_0^{-1/2}\mathbf{S}_k\mathbf{S}_0^{-1/2}$.

The SOBI estimate is computed in practice using the whitening. The most popular algorithm for the approximate joint diagonalization of several matrices is based on Jacobi rotations (Clarkson 1988). The algorithm (25.6) in Sect. 25.2.3 was recently introduced in Miettinen et al. (2014b), where the statistical properties of the SOBI estimate were studied.

We use the notation $\mathbf{S_1}, \ldots, \mathbf{S}_K$, even though the set of lags can be something else than the K first. The separation performance depends on which lags are included, and different sets of lags are appropriate for different kind of data. Only case-specific guidelines for the selection of lags have been presented, see, for example, Tang et al. (2005).

In addition to the original SOBI, some modified versions have been proposed. The covariance matrix $\mathbf{S}_0$ has a special role in SOBI as it is exactly diagonalized. This unequal treatment can be removed by the use of non-orthogonal joint diagonalization algorithms, which replace the orthogonality constraint by some different ones, see, for example, Yeredor (2002); Yeredor et al. (2004); Ziehe et al. (2004).

In the deflation-based SOBI (Miettinen et al. 2014a) method, the rows of the matrix $\mathbf{U}$ are found one by one, and the jth row of $\mathbf{U}$ maximizes

$$\sum_{k=1}^{K} (\mathbf{u}_j'\mathbf{S}_k\mathbf{u}_j)^2, \tag{25.3}$$

under the constraints $\mathbf{u}_i'\mathbf{u}_j = \delta_{ij}$, $i = 1, \ldots, j$, where δ_{ij} is the Kronecker delta. In Miettinen et al. (2014b) the deflation-based SOBI estimator was compared with the original, symmetric, SOBI estimator. The conclusion was that the symmetric version is the preferable choice. Miettinen et al. (2014a) also introduced a family of functionals where criterion (25.3) is replaced by

$$\sum_{k=1}^{K} G(\mathbf{u}_j'\mathbf{R}_k\mathbf{u}_j),$$

with continuously differentiable functions G. It was demonstrated that different choices of G may yield considerably more efficient estimates than (25.3). The rest of the paper is devoted to a study on the use of alternative diagonality criteria with the symmetric SOBI.

25.2.3 Other Diagonality Criteria

The idea of SOBI is that it searches for the unmixing matrix that makes selected autocovariance matrices as diagonal as possible. To measure the diagonality by the sum of squares of the diagonal elements as in (25.2) is natural because under the orthogonality constraint, maximizing (25.2) is equivalent to minimizing the sum of squares of the off-diagonal elements

$$\sum_{k=1}^{K} ||\text{off}(\mathbf{U}\mathbf{R}_k\mathbf{U}')||^2,$$

where $\text{off}(\mathbf{A}) = \mathbf{A} - \text{diag}(\mathbf{A})$. This property does not hold when we switch (25.2) to

$$\sum_{k=1}^{K}\sum_{j=1}^{p} G(\mathbf{u}_j'\mathbf{R}_k\mathbf{u}_j), \tag{25.4}$$

where the function G is a continuously differentiable and even function with $G(0) = 0$, $g(x) = G'(x) \leq 0$ for $x < 0$ and $g(x) \geq 0$ for $x > 0$. In this context, only the function values in interval $[-1, 1]$ matter. We will focus on functions of the form

$$G_a(x) = |x|^a, \quad a > 1. \tag{25.5}$$

Let us denote by $\boldsymbol{\Gamma}^a = \mathbf{U}\mathbf{S}_0^{-1/2}$ the unmixing matrix functional for which $\mathbf{U}$ maximizes (25.4) with $G = G_a$. The functional $\boldsymbol{\Gamma}^a$ is affine equivariant. The larger the a is, the more weight is put to the matrices with large autocovariances. If each element of $\boldsymbol{\Lambda}_k$ is greater than the corresponding elements of $\boldsymbol{\Lambda}_j$, $j \neq k$, then $\boldsymbol{\Gamma}^a$ converges to the AMUSE estimate $\boldsymbol{\Gamma}_k$ as $a \to \infty$.

The estimating equations for $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)'$ are found using Lagrange multipliers method similarly as in Miettinen et al. (2014b). The equations are

$$\mathbf{u}_i'\mathbf{T}(\mathbf{u}_j) = \mathbf{u}_j'\mathbf{T}(\mathbf{u}_i) \text{ and } \mathbf{u}_i'\mathbf{u}_j = \delta_{ij}, \quad i, j = 1, \ldots, p,$$

where $\mathbf{T}(\mathbf{u}) = \sum_{k=1}^{K} g_a(\mathbf{u}'\mathbf{R}_k\mathbf{u})\mathbf{R}_k\mathbf{u}$.

The estimating equations gave the idea to an algorithm with the two steps

$$\begin{aligned} &\text{step1}: \quad \mathbf{T} \leftarrow (\mathbf{T}(\mathbf{u}_1), \ldots, \mathbf{T}(\mathbf{u}_p))' \\ &\text{step2}: \quad \mathbf{U} \leftarrow (\mathbf{T}\mathbf{T}')^{-1/2}\mathbf{T}. \end{aligned} \tag{25.6}$$

The algorithm is easily adjusted to solve the joint diagonalization problem for any G. For $a = 2$, algorithm (25.6) and the algorithm based on Jacobi rotations yield equal solutions. Algorithm (25.6) is slower when the components are difficult to

separate, that is, when the sample size is small or there are mutually almost similar components.

25.3 Asymptotical Properties

The theorems of this section are straightforward generalizations of the corresponding theorems for the standard symmetric SOBI estimate. To get the idea of the proofs, see Miettinen et al. (2014b).

25.3.1 General Case

Since the population autocovariance matrices are symmetric in the BSS model (25.1), we estimate them using the symmetrized sample autocovariance matrices

$$\hat{\mathbf{S}}_k = \frac{1}{2(T-k)} \sum_{t=1}^{T-k} (\mathbf{x}_t \mathbf{x}_{t+k}' + \mathbf{x}_{t+k} \mathbf{x}_t'), \quad k = 0, 1, 2, \ldots$$

To guarantee the identifiability of the unmixing matrix with the autocovariance matrices $\hat{\mathbf{S}}_0, \hat{\mathbf{S}}_1, \ldots, \hat{\mathbf{S}}_K$, $K \geq 2$, we assume

(A4) the diagonal elements of $\sum_{k=1}^{K} G_a(\boldsymbol{\Lambda}_k)$ are strictly decreasing.

Also the order of the component time series is fixed by Assumption (A4). For the next Theorem we assume root-T consistency of the sample autocovariance matrices $\hat{\mathbf{S}}_0, \hat{\mathbf{S}}_1, \ldots, \hat{\mathbf{S}}_K$.

(A5) $\boldsymbol{\Omega} = \mathbf{I}_p$ and $\sqrt{T}(\hat{\mathbf{S}}_k - \boldsymbol{\Lambda}_k) = O_p(1), k = 0, 1, \ldots, K$ as $T \to \infty$.

The assumption $\boldsymbol{\Omega} = \mathbf{I}_p$ is made without loss of generality because the affine equivariance of the estimate implies that $(\hat{\boldsymbol{\Gamma}} \boldsymbol{\Omega} - \mathbf{I}_p)$ does not depend on $\boldsymbol{\Omega}$, and since $\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma} = (\hat{\boldsymbol{\Gamma}} \boldsymbol{\Omega} - \mathbf{I}_p)\boldsymbol{\Gamma}$, the limiting distribution of $\sqrt{T}\,\mathrm{vec}(\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})$ is derived easily from that of $\sqrt{T}\,\mathrm{vec}(\hat{\boldsymbol{\Gamma}} \boldsymbol{\Omega} - \mathbf{I}_p)$.

Theorem 25.1 *Under the assumptions (A1)–(A5) we have, for* $a > 1$, $\hat{\boldsymbol{\Gamma}}^a = (\hat{\boldsymbol{\gamma}}_1^a, \ldots, \hat{\boldsymbol{\gamma}}_p^a)' \to_p I_p$, *and for* $j = 1, \ldots, p$,

$$\sqrt{T}\hat{\gamma}_{jj}^a = -\frac{1}{2}\sqrt{T}((\hat{\mathbf{S}}_0)_{jj} - 1) + o_p(1), \quad i = j$$

$$\sqrt{T}\hat{\gamma}_{ji}^a = \frac{\sum_k (g_a(\lambda_{kj}) - g_a(\lambda_{ki}))\left[(\sqrt{T}\hat{\mathbf{S}}_k)_{ji} - \lambda_{kj}(\sqrt{T}\hat{\mathbf{S}}_0)_{ji}\right]}{\sum_k (g_a(\lambda_{kj}) - g_a(\lambda_{ki}))(\lambda_{kj} - \lambda_{ki})} + o_p(1), \quad i \neq j.$$

Notice that, asymptotically, the diagonal elements of $\hat{\boldsymbol{\Gamma}}^a$ do not depend on $\hat{\mathbf{S}}_1, \ldots, \hat{\mathbf{S}}_K$, and consequently, they neither depend on the diagonality criterion. Therefore, inspecting the limiting variances of the off-diagonal elements is enough when comparing the different estimates. The off-diagonal element $\hat{\gamma}_{ij}$ depends asymptotically only on the ith and jth source component.

Limiting joint multivariate normality of the unmixing matrix estimate would follow from the joint limiting multivariate normality of the sample autocovariance matrices.

Corollary 25.1 *Under the assumptions (A1)–(A5), if the joint limiting distribution of*

$$\sqrt{T}\left[\text{vec}(\hat{\mathbf{S}}_0, \hat{\mathbf{S}}_1, \ldots, \hat{\mathbf{S}}_K) - vec(\mathbf{I}_p, \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_K)\right]$$

is a (singular) $(K+1)p^2$-variate normal distribution with mean value zero, then the joint limiting distribution of $\sqrt{T}\text{vec}(\hat{\boldsymbol{\Gamma}}^a - \boldsymbol{\Gamma}^a)$ is a singular p^2-variate normal distribution.

25.3.2 Linear Process Model

The limiting variances of the elements of $\hat{\boldsymbol{\Gamma}}$ can be calculated when the components of $\mathbf{z}_t$ in the BSS model (25.1) are uncorrelated linear processes, that is,

$$\mathbf{z}_t = \sum_{j=-\infty}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\epsilon}_{t-j}, \tag{25.7}$$

where $\boldsymbol{\Psi}_j, j = 0, \pm 1, \pm 2, \ldots$, are diagonal matrices with diagonal elements $\psi_{j1}, \ldots, \psi_{jp}$ satisfying $\sum_{j=-\infty}^{\infty} \boldsymbol{\Psi}_j^2 = \mathbf{I}_p$, and the innovations $\boldsymbol{\epsilon}_t$ are iid p-variate random vectors with $E(\boldsymbol{\epsilon}_t) = \mathbf{0}$ and $\mathbf{C}ov(\boldsymbol{\epsilon}_t) = \mathbf{I}_p$. Also $\mathbf{x}_t = \boldsymbol{\Omega}\mathbf{z}_t$ is then a multivariate linear process. Note that linear processes cover a very wide class of second-order stationary processes and that, for example, all invertible ARMA(p, q) processes are linear processes. See Chap. 3 in Brockwell and Davis (1991).

We further assume that

(A6) the components of $\boldsymbol{\epsilon}_t$ have finite fourth order moments, and
(A7) the components of $\boldsymbol{\epsilon}_t$ are exchangeable and marginally symmetric.

Assumption (A7) implies that all third moments of $\boldsymbol{\epsilon}_t$ are zero. In the following we write $\beta_{ij} = E(\epsilon_{ti}^2 \epsilon_{tj}^2)$, $i, j = 1, \ldots, p$. Notice that if (A5) is replaced by the assumption that the components of $\boldsymbol{\epsilon}_t$ are mutually independent, then, in this independent component model case, $\beta_{ij} = 1$ for $i \neq j$. If one further assumes that innovations $\boldsymbol{\epsilon}_t$ are iid from $N_p(\mathbf{0}, \mathbf{I}_p)$, then $\beta_{ii} = 3$ and $\beta_{ij} = 1$ for all $i \neq j$.

In the linear process model, the joint limiting distribution of the sample autocovariance matrices $\hat{\mathbf{S}}_k^S$, $k = 1, \ldots, K$ is multivariate normal (Su and Lund 2012; Miettinen et al. 2012). First define

$$\mathbf{F}_k = \sum_{t=-\infty}^{\infty} \boldsymbol{\psi}_t \boldsymbol{\psi}_{t+k}', \quad k = 0, 1, 2, \ldots$$

where the diagonal elements of $\mathbf{F}_k$ are the autocovariances at lag k of the components of $\mathbf{z}_t$. The limiting covariances of the elements of $\hat{\mathbf{S}}_l^S$ and $\hat{\mathbf{S}}_m^S$ are given in the $p \times p$ matrix $\mathbf{D}_{lm}$ with elements

$$(\mathbf{D}_{lm})_{ii} = (\beta_{ii} - 3)(\mathbf{F}_l)_{ii}(\mathbf{F}_m)_{ii} + \sum_{k=-\infty}^{\infty} ((\mathbf{F}_{k+l})_{ii}(\mathbf{F}_{k+m})_{ii} + (\mathbf{F}_{k+l})_{ii}(\mathbf{F}_{k-m})_{ii}),$$

$$(\mathbf{D}_{lm})_{ij} = \frac{1}{2} \sum_{k=-\infty}^{\infty} ((\mathbf{F}_{k+l-m})_{ii}(\mathbf{F}_k)_{jj} + (\mathbf{F}_k)_{ii}(\mathbf{F}_{k+l-m})_{jj}) + \frac{1}{4}(\beta_{ij} - 1)(\mathbf{F}_l + \mathbf{F}_l')_{ij}(\mathbf{F}_m + \mathbf{F}_m')_{ij}, \quad i \neq j.$$

The limiting distributions of the rows of $\hat{\boldsymbol{\Gamma}}^a$ in the linear model are given in the following Theorem.

Theorem 25.2 *Assume that* $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ *is an observed time series from the linear process (25.7) that satisfies (A1)–(A7). Assume (wlog) that* $\boldsymbol{\Omega} = \mathbf{I}_p$. *If* $\hat{\boldsymbol{\Gamma}}^a = (\hat{\boldsymbol{\gamma}}_1^a, \ldots, \hat{\boldsymbol{\gamma}}_p^a)'$ *is the SOBI estimate with* $G = G_a$, *then the limiting distribution of* $\sqrt{T}(\hat{\boldsymbol{\gamma}}_j^a - \mathbf{e}_j)$ *is a p-variate normal distribution with mean zero and covariance matrix*

$$ASV(\hat{\boldsymbol{\gamma}}_j^a) = \sum_{r=1}^{p} ASV(\hat{\gamma}_{jr}^a)\mathbf{e}_r\mathbf{e}_r'$$

where, for $i \neq j$,

$$ASV(\hat{\gamma}_{jj}^a) = \frac{1}{4}(\mathbf{D}_{00})_{jj},$$

$$ASV(\hat{\gamma}_{ji}^a) = \frac{\sum_{l,m}(g_a(\lambda_{lj}) - g_a(\lambda_{li}))(g_a(\lambda_{mj}) - g_a(\lambda_{mi}))(\mathbf{D}_{lm})_{ji}}{\left[\sum_k (g_a(\lambda_{kj}) - g_a(\lambda_{ki}))(\lambda_{kj} - \lambda_{ki})\right]^2} + \frac{-2v_{ji}\sum_k (g_a(\lambda_{kj}) - g_a(\lambda_{ki}))(\mathbf{D}_{k0})_{ji} + v_{ji}^2(\mathbf{D}_{00})_{ji}}{\left[\sum_k (g_a(\lambda_{kj}) - g_a(\lambda_{ki}))(\lambda_{kj} - \lambda_{ki})\right]^2},$$

with $v_{ji} = \sum_k (g_a(\lambda_{kj}) - g_a(\lambda_{ki}))\lambda_{kj}$.

25.3.3 Adaptive SOBI Estimator Based on the Asymptotic Theory

The results of Theorem 25.2 can be employed similarly as in the adaptive deflation-based FastICA (Miettinen et al. 2014a), where, for each component, the asymptotically most efficient nonlinearity function is chosen from a set of candidate nonlinearities. Let $\mathbf{a}$ denote the candidate set of parameter values of 25.5. In the simulation study we will have $\mathbf{a} = \{1.5, 2, 3, 4\}$. When it comes to computation times, having a larger set would not be a problem. The structure of the adaptive SOBI estimator is the following.

step 1 Find estimates of the source components, for example using the regular SOBI.
step 2 Estimate the asymptotic variances of Theorem 25.2, and find $a_0 \in \mathbf{a}$ which minimizes the sum of the variances.
step 3 Estimate the unmixing matrix using SOBI with a_0.

The same arguments as in the proof of Theorem 3 in Miettinen et al. (2014a), state that the adaptive estimate is asymptotically equivalent with $\hat{\boldsymbol{\Gamma}}^{a_0}$, where a_0 minimizes the sum of the asymptotic variances in $\mathbf{a}$.

The asymptotic variances can be estimated from the data at hand, when the source time series are linear processes. If we further assume that the innovations are normally distributed, we only need to compute sample autocovariances from the estimated source time series. Otherwise, the MA coefficients have to be estimated, so that the estimation of β_{ij}, $i,j = 1,\ldots,p$ is possible. Since we are now more interested in the order of sums of variances for different values of a than the exact values of the variances, it is reasonable to set $\beta_{ii} = 3$ and $\beta_{ij} = 1$, $i \neq j$ as if the innovations were normal.

To approximate the infinite series in $\mathbf{D}_{lm}$, one needs to include all non-zero autocovariances. On the other hand, the length of the time series should be long enough so that the estimation of all autocovariances is reliable. In the simulation study we will compute the 50 first autocovariances.

Functions to compute the asymptotic variances are provided in the R package BSSasymp (Miettinen et al. 2013).

25.4 Efficiency Study

In this section we study the impact of the new diagonality measures to asymptotic and finite-sample efficiency in estimation of uncorrelated linear process time series.

25.4.1 Minimum Distance Index

As the unmixing matrix is identifiable only up to sign changes, rescaling and permutation of the rows, it is satisfactory that $\hat{\boldsymbol{\Gamma}}\boldsymbol{\Omega}$ converges to some $C \in \mathscr{C}$, where

$$\mathscr{C} = \{\mathbf{C} \ : \ \text{each row and column of } \mathbf{C} \text{ has exactly one non-zero element}\}.$$

Hence, the accuracy of the estimate $\hat{\boldsymbol{\Gamma}}$ in a simulation study is often measured as a distance from $\hat{\boldsymbol{\Gamma}}\boldsymbol{\Omega}$ to $\mathscr{C}$. For an overview of performance indices, see, for example, Nordhausen et al. (2011a). The minimum distance index (MDI) (Ilmonen et al. 2010) is defined by

$$\hat{D} = D(\hat{\boldsymbol{\Gamma}}\boldsymbol{\Omega}) = \frac{1}{\sqrt{p-1}} \inf_{\mathbf{C}\in\mathscr{C}} \|\mathbf{C}\hat{\boldsymbol{\Gamma}}\boldsymbol{\Omega} - \mathbf{I}_p\|.$$

MDI is invariant with respect to the change of the mixing matrix, and it is scaled so that $0 \leq \hat{D} \leq 1$. The smaller the MDI-value, the better is the performance.

The MDI is an unquestionable choice, when the estimate $\hat{\boldsymbol{\Gamma}}$ is asymptotically normal, because it has the following property: if $\sqrt{T}\,\mathrm{vec}(\hat{\boldsymbol{\Gamma}}\boldsymbol{\Omega} - \mathbf{I}_p) \to N_{p^2}(\mathbf{0}, \boldsymbol{\Sigma})$, the limiting distribution of $T(p-1)\hat{D}^2$ is that of a weighted sum of independent chi squared variables with the expected value

$$\mathrm{tr}\left((\mathbf{I}_{p^2} - \mathbf{D}_{p,p})\boldsymbol{\Sigma}\,(\mathbf{I}_{p^2} - \mathbf{D}_{p,p})\right).$$

The expected value thus equals the sum of the limiting variances of the off-diagonal elements of $\sqrt{T}\,\mathrm{vec}(\hat{\boldsymbol{\Gamma}}\boldsymbol{\Omega} - \mathbf{I}_p)$.

25.4.2 Simulation Setups

To examine some finite-sample properties of the estimates obtained using different values of a in (25.5) and with the adaptive choice of a, two sets of time series processes are chosen to comprise the source vectors of the BSS model. The source time series in the two models are

(i) four AR(4) time series with coefficient vectors (0.6, 0, −0.2), (0.4, 0.3, −0.2), (0, −0.2, 0.3) and (−0.2, −0.2, −0.2), respectively, and normal innovations,
(ii) three MA(5) time series with coefficient vectors (0.4, 0.3, 0.1, 0.6, 0.1), (0.2, −0.3, −1.2, −0.2, −0.2) and (0.7, 0.2, 0.1, 0.6, −0.9), respectively, and normal innovations.

For each length of time series, 10,000 repetitions were carried out and the MDI-values were computed.

In the third setup, the aim is to conduct a wide study on the asymptotic efficiencies by generating randomly 10,000 pairs of ARMA processes. For each pair we computed the sum of the asymptotic variances of the off-diagonal elements of $\hat{\boldsymbol{\Gamma}}\boldsymbol{\Omega}$ for $a = 1.1, 1.2, \ldots, 4$. The ARMA processes were created so that first the AR and MA orders were sampled (uniform distribution) from $\{0, 1, 2, 3, 4\}$ and $\{0, \ldots, 10\}$, respectively. Then each AR and MA coefficient was generated from the uniform distribution with bounds -0.8 and 0.8. To guarantee the stationarity of the time series, the AR coefficients had the constraint that the sum of their absolute value is less than 1.

The set of lags $\{1, \ldots, 10\}$ is used, except in model (ii) where the idea is to investigate the impact of including too many matrices. There $\{1, \ldots, 5\}$ and $\{1, \ldots, 20\}$ are used.

25.4.3 Simulation Results

Figure 25.1 presents the simulation results from model (i), which is an example of a case where some other choice than $a = 2$ yields better results, even if the asymptotically optimal set of lags (here $\{1, 2, 3\}$) is used. In model (i) for lags $\{1, \ldots, 10\}$, the optimal value of a is 3.58 and then the expected value of $T(p-1)\hat{D}^2$ is 23.61. The corresponding values for $a = 2$ and sets of lags $\{1, 2, 3\}$ and

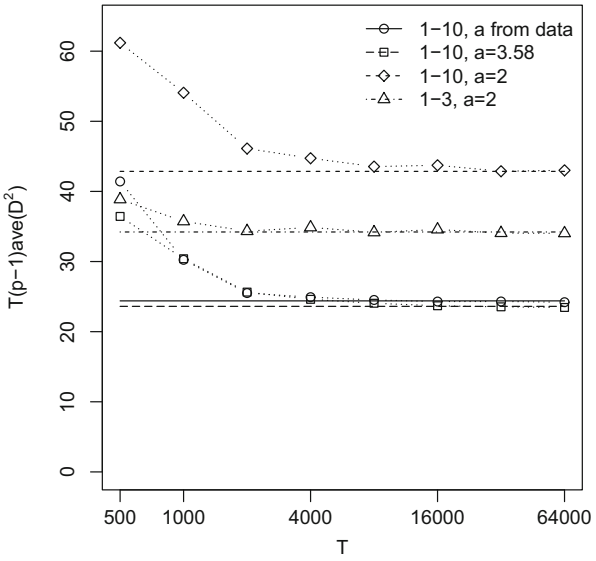

Fig. 25.1 The averages of $T(p-1)\hat{D}^2$ over 10,000 repetitions of the observed time series with length T from model (i) for the different estimates. The *horizontal lines* give the expected values of the limiting distributions of $T(p-1)\hat{D}^2$

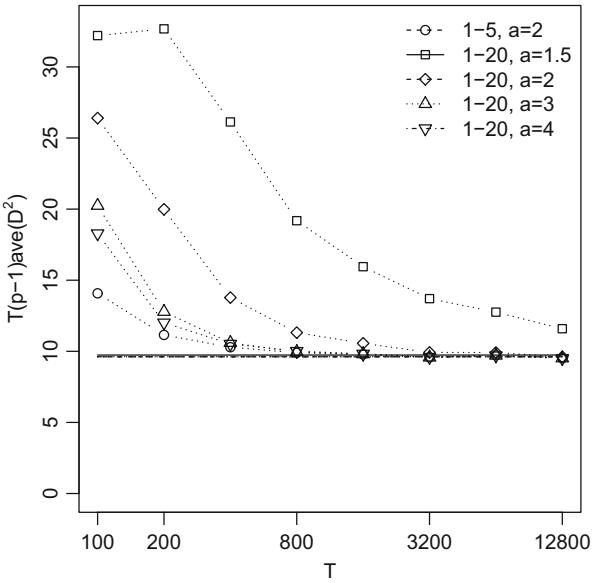

Fig. 25.2 The averages of $T(p-1)\hat{D}^2$ over 10,000 repetitions of the observed time series with length T from model (ii) for the different estimates. The *horizontal lines* give the expected values of the limiting distributions of $T(p-1)\hat{D}^2$

$\{1,\ldots,10\}$ are 34.22 and 42.86, respectively. The fourth estimator in this simulation setup is the estimator defined in 25.3.3. Its expected value for $T(p-1)\hat{D}^2$ is that of $a=4$, which is 24.39.

In model (ii) the expected values of $T(p-1)\hat{D}^2$ are very close together for estimates with different diagonality criteria. For $a=1.5,2,3,4$ they are 9.74, 9.62, 9.62, and 9.61, respectively, when lags 1, 2, 3, 4, and 5 are included. Notice that including autocovariance matrices whose all expected values are zeros, which is the case here for lags greater than 5, has no effect on the asymptotic variances. With small sample sizes, however, having too many matrices decreases the efficiency. Figure 25.2 shows how the estimators perform when the set of lags is $\{1,\ldots,20\}$. For comparison, the results of the estimator with lags $\{1,\ldots,5\}$ and $a=2$ are plotted in the same figure. The results of the estimators with lags $\{1,\ldots,5\}$ and $a=1.5,3,4$ were very similar to those of $a=2$, but, to keep the figure clear, they are not plotted. The impact of extra matrices is rather small for $a=3$ and $a=4$, but $a=2$ and especially $a=1.5$ suffer when the sample size is small.

Figure 25.3 plots the averages and the quartiles of $ASV(\hat{\gamma}_{12})+ASV(\hat{\gamma}_{21})$ for $a=1.1,1.2,\ldots,4$ from 10,000 pairs of randomly generated source time series processes. Based on the first two quartiles, all estimators seem to be asymptotically equally efficient. The average line shows that small values of a yield occasionally much larger variances. For $a\geq 2$ the estimators are fairly even. It means that in

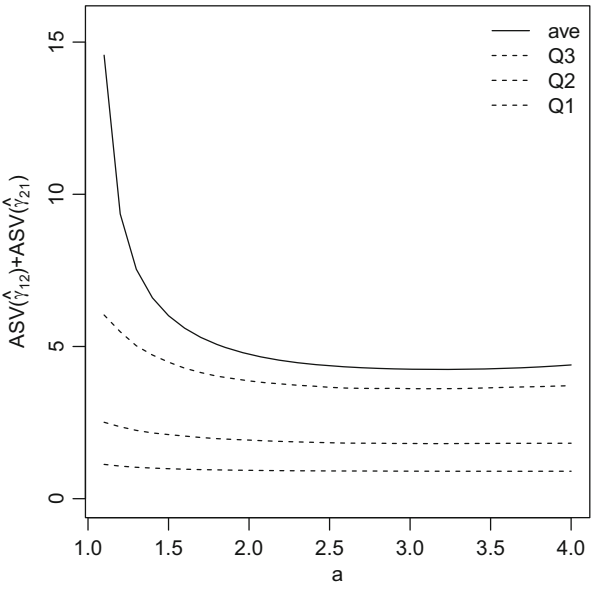

Fig. 25.3 The averages and the quartiles of the sum $ASV(\hat{\gamma}_{12}) + ASV(\hat{\gamma}_{21})$ from 10,000 pairs of source time series, that were generated as described in Sect. 25.4.2

Table 25.1 Distribution of the ratio between $ASV(\hat{\gamma}_{12}) + ASV(\hat{\gamma}_{21})$ for the optimal a and for that of $a = 2$

Range of a	$(0, 0.2]$	$(0.2, 0.4]$	$(0.4, 0.6]$	$(0.6, 0.8]$	$(0.8, 0.9]$	$(0.9, 1]$
$1.1, \ldots, 4$	8	63	252	1229	2130	6318
$2, 3, 4$	8	57	240	1133	1957	6572

high-dimensional cases, the choice of the diagonality criterion has probably only a minor effect.

Still utilizing the same 10,000 pairs of ARMA processes, we next take a look what kind of efficiency gain can be achieved when $p = 2$. In Table 25.1 we have listed frequencies of occurrences that the ratio between $ASV(\hat{\gamma}_{12}) + ASV(\hat{\gamma}_{21})$ for the optimal a and for that of $a = 2$ falls into intervals $(0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$, $(0.8, 0.9]$ and $(0.9, 1]$. The upper line gives the frequencies, when the optimal value is searched over $a = 1.1, 1.2, \ldots, 4$, and the lower line, when only $a = 2, 3, 4$ are considered. The fact that the two lines are quite similar implies that values $a < 2$ are rarely significantly better than others, and that $ASV(\hat{\gamma}_{12}) + ASV(\hat{\gamma}_{21})$ as a function of a is usually smooth. In about 35 % of the cases the efficiency gain is more than 10 %, and at highest it was 91 %.

25.5 Discussion

In this paper we generalized the asymptotic analysis of the symmetric SOBI estimator given in Miettinen et al. (2014b), to cover other criteria for diagonality of a set of matrices than the standard least squares criterion. The results were utilized when constructing a new estimator which chooses asymptotically the most efficient criterion based on an initial estimate of the source time series. The new method is more efficient than the regular SOBI, but, on the other hand, it is computationally heavier. For consideration whether the longer computation time is worth taking, one should include the sample size and the dimension of the data. The sample size should be sufficiently large, say $T > 500$, so that the approximation of the variances pays off. If the number of the latent components is large, then different diagonality criteria yield probably almost equally good estimators.

The choice of the diagonality criterion relates to the choice of lags. In most cases where the regular ($a = 2$) SOBI is not optimal for the selected set of lags, one can achieve improvement also by picking a better set of lags. But as we showed in Sect. 25.4.3, sometimes even the optimal set of lags cannot obtain the efficiency of the optimal diagonality criterion. Future work could include studying the choice of lags based on data and how to combine it with the choice of the diagonality criterion.

Acknowledgements This research was supported by the Academy of Finland (grant 256291).

References

Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., Moulines, E.: A blind source separation technique using second-order statistics. IEEE Trans. Signal Process. **45**, 434–444 (1997)

Brockwell, P.J., Davis, R.A.: Time Series: Theory and Methods, 2nd edn. Springer, New York (1991)

Clarkson, D.B.: A least squares version of algorithm AS 211: the F-G diagonalization algorithm. Appl. Stat. **37**, 317–321 (1988)

Comon, P.: Independent component analysis - a new concept? Signal Process. **36**, 287–314 (1994)

Ilmonen, P., Nordhausen, K., Oja, H., Ollila, E.: A new performance index for ICA: properties computation and asymptotic analysis. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) Latent Variable Analysis and Signal Separation, pp. 229–236. Springer, Heidelberg (2010)

Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S.: Statistical properties of a blind source separation estimator for stationary time series. Stat. Probab. Lett. **82**, 1865–1873 (2012)

Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S.: BSSasymp: covariance matrices of some BSS mixing and unmixing matrix estimates (2013). R package version 1.0-0. http://cran.r-project.org/web/packages/BSSasymp

Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S.: Deflation-based separation of uncorrelated stationary time series. J. Multivar. Anal. **123**, 214–227 (2014)

Miettinen, J., Illner, K. Nordhausen, K., Oja, H., Taskinen, S., Theis, F.: Separation of uncorrelated stationary time series using autocovariance matrices (2014). http://arxiv.org/abs/1405.3388

Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S.: Deflation-based FastICA with adaptive choices of nonlinearities. IEEE Trans. Signal Process. **62**(21), 5716–5724 (2014)

Nordhausen, K., Ollila, E., Oja, H.: On the performance indices of ICA and blind source separation. In: Proceedings of 2011 IEEE 12th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2011), pp. 486–490 (2011)

Su, N., Lund, R.: Multivariate versions of Bartlett's formula. J. Multivar. Anal. **105**, 18–31 (2012)

Tang, A.C., Liu, J.-Y., Sutherland, M.T.: Recovery of correlated neuronal sources from EEG: the good and bad ways of using SOBI. NeuroImage **7**, 507–519 (2005)

Tong, L., Soon, V.C., Huang, Y.F., Liu, R.: AMUSE: a new blind identification algorithm. In: Proceedings of IEEE International Symposium on Circuits and Systems, pp. 1784–1787 (1990)

Yeredor, A.: Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. IEEE Trans. Signal Process. **50**(7) 1545–1553 (2002)

Yeredor, A., Ziehe, A., Müller, K.R.: Approximate joint diagonalization using a natural gradient approach. In: Independent Component Analysis and Blind Signal Separation, pp. 89–96. Springer, Berlin, Heidelberg (2004)

Ziehe, A., Müller, K.-R.: TDSEP - an efficient algorithm for blind separation using time structure. In: Proceedings of ICANN, pp. 675–680 (1998)

Ziehe, A., Laskov, P., Nolte, G., Müller, K.R.: A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. J. Mach. Learn. Res. **5**, 777–800 (2004)

Chapter 26
Robust Simultaneous Sparse Approximation

Esa Ollila

Abstract This chapter considers sparse underdetermined (ill-posed) multivariate multiple linear regression model known in signal processing literature as multiple measurement vector (MMV) model. The objective is to find good recovery of jointly sparse unknown signal vectors from the given multiple measurement vectors which are different linear combinations of the same known elementary vectors. The MMV model is an extension of the compressed sensing (CS) which is an emerging field that has attracted considerable research interest over the past few years. Recently, many popular greedy pursuit algorithms have been extended to MMV setting. All these methods, such as simultaneous normalized iterative hard thresholding (SNIHT), are not resistant to outliers or heavy-tailed errors. In this chapter, we develop a robust SNIHT method that computes the estimates of the sparse signal matrix and the scale of the error distribution simultaneously. The method is based on Huber's criterion and hence referred to as HUB-SNIHT algorithm. The method can be tuned to have a negligible performance loss compared to SNIHT under Gaussian noise, but obtains superior joint sparse recovery under heavy-tailed non-Gaussian noise conditions.

Keywords Compressed sensing • Multiple measurement vector model • Sparse multivariate regression • Iterative hard thresholding

26.1 Introduction

26.1.1 The MMV Model

The *compressed sensing (CS)* and *sparse signal reconstruction* (SSR) is a signal processing technique that exploits the fact that acquired data can have a sparse representation in some basis. It allows for solving ill-posed linear systems. The problem can be formulated as follows; see Elad (2010); Donoho (2006); Candes

E. Ollila (✉)
Department of Signal Processing and Acoustics, Aalto University, 02150 Espoo, Finland
e-mail: esa.ollila@aalto.fi

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_26

and Wakin (2008). Let $\mathbf{y} = (y_1, \ldots, y_M)^\top$ denote the observed data vector (*measurements*) modelled as

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{x} + \mathbf{e} \tag{26.1}$$

$$\text{i.e., } y_i = \boldsymbol{\phi}_{(i)}^\top \mathbf{x} + e_i, \quad i = 1, \ldots, M \tag{26.2}$$

where

$$\boldsymbol{\Phi} = \left(\boldsymbol{\phi}_1 \cdots \boldsymbol{\phi}_N\right) = \left(\boldsymbol{\phi}_{(1)} \cdots \boldsymbol{\phi}_{(M)}\right)^\top$$

is an $M \times N$ *measurement matrix*, also called the *design matrix*, which typically can have more column vectors $\boldsymbol{\phi}_i$ than row vectors $\boldsymbol{\phi}_{(j)}$ (so $M > N$ is not assumed), $\mathbf{x} = (x_1, \ldots, x_N)^\top$ is the unobserved *signal vector* and $\mathbf{e} = (e_1, \ldots, e_M)^\top$ is the (unobserved) random noise vector. The goal is to recover the signal vector $\mathbf{x}$ which is assumed to be *K-sparse* (i.e., it has $K < M$ *non-zero* elements) or *compressible* (i.e., it has a representation whose entries decay rapidly when sorted in order of decreasing magnitude). In some applications, the measurement matrix is also called the dictionary and its columns are referred to as atoms. In statistics, model (26.1) is called the multiple linear regression model and $\mathbf{x}$ the vector of regression coefficients. Note that compressible signals are well approximated by K-sparse signals and typically in many applications $K \ll N$. In the noiseless case and when $\mathbf{x}$ is K-sparse, finding the exact solution is a subset selection problem, which is an NP-hard combinatorial optimization problem. However, when the measurement matrix $\boldsymbol{\Phi}$ satisfies certain coherence conditions, several reconstruction algorithms are guaranteed to give exact recovery of $\mathbf{x}$ in polynomial time. Furthermore, when the measurements are corrupted by noise with bounded norm and/or if $\mathbf{x}$ is not exactly K-sparse, bounds on the recovery error are known.

A natural extension of the CS model is the *multiple measurement vector (MMV) model* where a single measurement matrix is utilized to obtain multiple measurement vectors, i.e.,

$$\mathbf{y}_i = \boldsymbol{\Phi}\mathbf{x}_i + \mathbf{e}_i, \quad i = 1, \ldots, Q$$

and the goal is to recover the set of sparse/compressible unknown signal vectors $\mathbf{x}_i$, $i = 1, \ldots, Q$. In matrix form, the MMV model can be written as

$$\mathbf{Y} = \boldsymbol{\Phi}\mathbf{X} + \mathbf{E},$$

where $\mathbf{Y} = \left(\mathbf{y}_1 \cdots \mathbf{y}_Q\right) \in \mathbb{R}^{M \times Q}$, $\mathbf{X} = \left(\mathbf{x}_1 \cdots \mathbf{x}_Q\right) \in \mathbb{R}^{N \times Q}$ and $\mathbf{E} = \left(\mathbf{e}_1 \cdots \mathbf{e}_Q\right) \in \mathbb{R}^{M \times Q}$ collect the measurement, the signal and the error vectors, respectively. When only a single measurement vector ($Q = 1$) is available, the MMV model reduces to the standard CS/SSR model (26.1), also referred to as single measurement vector (SMV) model. Then, rather than recovering the sparse/compressible target signals $\mathbf{x}_i$

separately (one-by-one) using standard CS reconstruction algorithms, one attempts to simultaneously (jointly) recover all signals. The key assumption for the success of this scheme is that *the nonzero values occur on a common location set*. In other words, the signal matrix $\mathbf{X}$ is assumed to be *K-rowsparse*, i.e., at most K rows of $\mathbf{X}$ contain non-zero entries. This means that the *row-support* of $\mathbf{X}$, defined as the index set of rows containing non-zero elements

$$\operatorname{rsupp}(\mathbf{X}) = \{i \in \{1, \ldots, N\} \,:\, x_{ij} \neq 0 \text{ for some } j\},$$

has cardinality less or equal to K. This is denoted shortly as $\|\mathbf{X}\|_0 = |\operatorname{rsupp}(\mathbf{X})| \leq K$, where $\|\cdot\|_0$, called the row-ℓ_0 quasi-norm, counts the number of nonzero rows of its matrix-valued argument.

Joint estimation offers both computational benefits as well as increased reconstruction accuracy. Obviously one could apply any CS method to recover each $\mathbf{x}_i$ separately (one-by-one) given the measurement vector $\mathbf{y}_i$, the measurement matrix $\boldsymbol{\Phi}$ and the sparsity level K. Nevertheless, since the vectors all share a common support, one can improve upon the recovery accuracy by exploiting the joint information (Tropp et al. 2006; Tropp 2006; Chen and Huo 2006; Gribonval et al. 2008; Eldar and Rauhut 2010; Duarte and Eldar 2011; Blanchard et al. 2014). The objective of simultaneous sparse approximation, also referred to as *multichannel sparse recovery* (Eldar and Rauhut 2010), can thus be stated as follows:

find a row sparse approximation of the signal matrix $\mathbf{X}$ *based on the MMV measurement matrix* $\mathbf{Y}$*, knowing only the measurement matrix* $\boldsymbol{\Phi}$ *and the sparsity level* K*.*

Simultaneous sparse approximation problem occurs when one has multiple observations (multichannel recordings) of the latent sparse signal contaminated by additive noise. Then the multiple measurements can provide a better estimate of the underlying sparse signal matrix. Such situations arise in various fields including electroencephalography and magnetoencephalography (EEG/MEG) (Gorodnitsky et al. 1995; Ou et al. 2009), equalization of sparse communications channels (Cotter and Rao 2002), blind source separation (Gribonval and Zibulevsky 2010), and direction-of-arrival estimation of sources in array processing (Malioutov et al. 2005).

26.1.2 Contributions and Outline

Many of the greedy pursuit reconstruction algorithms developed for SMV model such as orthogonal matching pursuit (OMP), normalized iterative hard thresholding (NIHT), Compressive Matching Pursuit (CoSaMP) have been extended for solving MMV problems in Tropp et al. (2006); Chen and Huo (2006); Blanchard et al. (2014). These methods are guaranteed to perform very well provided that suitable conditions (e.g., incoherence of $\boldsymbol{\Phi}$ and non-impulsive noise conditions) are met. It

is important to notice that the derived (worst case) recovery bounds depend linearly on the (Frobenius) norm of the error, $\|\mathbf{E}\|$, and thus the methods are not guaranteed to provide accurate reconstruction/approximation under heavy-tailed non-Gaussian noise.

Despite the vast interest on sparse approximation during the past decade, *sparse and robust methods* for MMV model have been considered in the literature only recently, for example, in Ollila (2015). In the SVM setting, however, several robustification of greedy CS algorithms exists such the Lorentzian IHT (Carrillo and Barner 2011) or robust versions of the CoSaMP and orthogonal matching pursuit (OMP) algorithms proposed in Razavi et al. (2012). In this chapter we extend the robust IHT algorithm developed in Ollila et al. (2014a,b) to MMV model. We utilize the criterion function proposed by Huber (1981, cf. Sects. 7.7 and 7.8) in the overdetermined ($M > N$) multiple linear regression model. (In the robust statistics literature, Huber's approach is also sometimes referred to as "Huber's approach 2".) It is important to note that standard robust loss functions require a preliminary robust estimate of scale. Obtaining such estimate is a very difficult problem in the ill-posed MMV model. Utilizing Huber's criterion allows us to estimate the scale simultaneously and hence avoids the problems caused by using a possibly corrupted and inaccurate preliminary estimate of scale.

Let us offer a brief outline of the chapter. First, in Sect. 26.2, we formulate a robust constrained objective function for the MMV problem and review standard robust loss functions that can be utilized in the criterion. Section 26.3 introduces the simultaneous normalized IHT (SNIHT) algorithm (Blanchard et al. 2014) as well as the new robust SNIHT method based on Huber's criterion, referred to as HUB-SNIHT. Section 26.4 provides a wealth of simulation examples illustrating the effectiveness of the HUB-SNIHT method in reconstructing a K-rowsparse signal matrix in various noise conditions and signal to noise (SNR) settings. We also illustrate that in low SNR setting and in heavy-tailed noise, accurate joint sparse recovery is possible by using the robust HUB-SNIHT whereas reconstruction based on a single channel information only fails, thus illustrating that joint sparse recovery is superior to applying standard sparse reconstruction methods to each channel individually. Section 26.5 concludes.

26.1.3 Matrix Preliminaries and Notations

For a matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ and an index set $\Gamma \subset \{1,\ldots,M\}$ of cardinality K, we denote by $\mathbf{A}_\Gamma$ (resp. $\mathbf{A}_{(\Gamma)}$) the $M \times K$ (resp. $K \times N$) matrix restricted to the columns (resp. rows) of $\mathbf{A}$ indexed by the set Γ. Notation $\mathbf{A}|_\Gamma$ refers to sparsified version of $\mathbf{A}$ such that the entries in the rows indexed by the set Γ remain unchanged while all other rows have all entries set to 0. Furthermore, if $f : \mathbb{R} \to \mathbb{R}$, then $f(\mathbf{A})$ refers to element-wise application of the function to its matrix valued argument, so $f(\mathbf{A}) \in \mathbb{R}^{M\times N}$ with $[f(\mathbf{A})]_{ij} = f(a_{ij})$. Let $H_K(\cdot)$ denote the *hard thresholding*

operator: for a matrix $\mathbf{X} \in \mathbb{R}^{N\times Q}$, $H_K(\mathbf{X})$ retains the elements of the K rows of $\mathbf{X}$ that possess largest ℓ_2-norms and set elements of the other rows to zero.

The usual Euclidean norm on vectors will be written as $\|\cdot\|$. The matrix space $\mathbb{R}^{M\times N}$ is equipped with the usual inner product

$$\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{Tr}(\mathbf{B}^\top \mathbf{A}) = \sum_{i=1}^{M}\sum_{j=1}^{N} a_{ij}b_{ij}$$

where the trace of a (square) matrix is the sum of its diagonal entries. We define the weighted inner product as

$$\langle \mathbf{A}, \mathbf{B}\rangle_{\mathbf{W}} = \sum_{i=1}^{M}\sum_{j=1}^{N} w_{ij}a_{ij}b_{ij}$$

where $\mathbf{W}$ is $M \times N$ matrix of positive weights. Note that $\langle \mathbf{A}, \mathbf{B}\rangle_{\mathbf{W}}$ reduces to conventional inner product when $\mathbf{W}$ is a matrix of ones. The Frobenius norm is given by the inner product as

$$\|\mathbf{A}\| = \sqrt{\langle \mathbf{A}, \mathbf{A}\rangle}$$

and $\|\mathbf{A}\|_{\mathbf{W}} = \sqrt{\langle \mathbf{A}, \mathbf{A}\rangle_{\mathbf{W}}}$ denotes the weighted Frobenius norm. Finally, the $\ell_{\infty,q}$ norm applies ℓ_∞ norm to rows and ℓ_q norm to the resulting vector, i.e., $\|\mathbf{A}\|_{\infty,q} = \left(\sum_{i=1}^{M}(\max_{j\in\{1,\ldots,N\}}|a_{ij}|)^q\right)^{1/q}$.

26.2 Optimization Problem of Multichannel Sparse Recovery

In this section we formulate the multichannel sparse recovery problem in the form of constrained optimization problem. First, in Sect. 26.2.1, we consider a general objective function based on a preliminary estimate of scale $\hat{\sigma}$ and a general loss function ρ. In Sect. 26.2.2 we review some standard loss functions commonly used in robust statistics literature (Huber 1981; Maronna et al. 2006) and discuss their properties. Since the scale parameter is unknown in practice, we introduce in Sect. 26.2.3 our constrained optimization problem that generalizes Huber's criterion for joint estimation of unknown location and scale parameters to the MMV model.

26.2.1 General Objective Function for Known Scale

Suppose that the error terms in $\mathbf{E}$ are i.i.d. random variables from a symmetric distribution with symmetry center at 0 and scale $\sigma > 0$. For a moment, we assume

that the scale σ is known or a preliminary estimate is available. In either case, let $\hat{\sigma}$ denote the known or estimated scale parameter. Under the above assumptions, reasonable approach for estimating K-rowsparse matrix $\mathbf{X}$ is to solve

$$\min_{\mathbf{X}\in\mathbb{R}^{N\times Q}} \sum_{j=1}^{Q}\sum_{i=1}^{M} \rho\left(\frac{y_{ij}-\boldsymbol{\phi}_{(i)}^{\top}\mathbf{x}_j}{\hat{\sigma}}\right) \quad \text{subject to} \quad \|\mathbf{X}\|_0 \le K, \tag{26.3}$$

where the loss function $\rho(e)$ is a continuous even function, nondecreasing in $e \ge 0$. For conventional least squares (LS) loss function $\rho(e) = e^2$, the scale $\hat{\sigma}$ factors out and the minimization problem becomes

$$\min_{\mathbf{X}} \|\mathbf{Y}-\boldsymbol{\Phi}\mathbf{X}\|^2 \quad \text{subject to} \quad \|\mathbf{X}\|_0 \le K.$$

The optimization problem (26.3) for any ρ is NP-hard. Therefore suboptimal approximation algorithms have been proposed which exploit the joint sparsity in different ways. One can divide these methods roughly into two main categories: (a) *convex optimization methods* and (b) *greedy pursuit approaches*.

Methods in the convex optimization class consider optimization problem in penalized (Lagrangian) form and replace the nonconvex ℓ_0-quasinorm of the signal matrix by the convex $\ell_{p,1}$ norm, which is known to enforce rowsparsity of the solution. In this case, the problem becomes multivariate LASSO regression (Tibshirani 1996) problem. For example, Turlach et al. (2005) and Tropp (2006) use $p = \infty$ and solve

$$\min_{\mathbf{X}} \|\mathbf{Y}-\boldsymbol{\Phi}\mathbf{X}\|^2 + \lambda\|\mathbf{X}\|_{\infty,1}, \tag{26.4}$$

where $\|\mathbf{X}\|_{\infty,1} = \sum_i \|\mathbf{x}_{(i)}\|_\infty$. Malioutov et al. (2005) use $p = 2$ and solve

$$\min_{\mathbf{X}} \|\mathbf{Y}-\boldsymbol{\Phi}\mathbf{X}\|^2 + \lambda\|\mathbf{X}\|_{2,1} \tag{26.5}$$

where $\|\mathbf{X}\|_{2,1} = \sum_i \|\mathbf{x}_{(i)}\|$. In (26.4) and (26.5), $\lambda > 0$ denotes a fixed penalty parameter that determines the degree of rowsparsity of the solution. Methods in the greedy pursuit class are iterative algorithms that are typically based on comparing correlations between the columns of $\boldsymbol{\Phi}$ with the current residual matrix and then updating the row-support based on this information. In this chapter, we employ greedy pursuit algorithm, called SNIHT, that is based on *projected gradient descent* algorithm. It is known to offer efficient and scalable solution for K-sparse approximation problem (Blumensath and Davies 2010). Another benefit of greedy pursuit approaches over the multivariate LASSO formulations are that they do not require a penalty parameter λ. Optimal selection of penalty parameter is a difficult task and can greatly affect the performance of the method. Here we use the information that signal matrix is K-rowsparse or is well approximated by being K-rowsparse (compressible).

The well-known problem with LS minimization used in (26.3) or in (26.4) and (26.5) is that it gives a very small weight on small residuals and a strong weight on large residuals, implying that even a single large outlier can have a large influence on the obtained result. For robustness, one should utilize a loss function that downweights large outliers. Such loss functions are reviewed next.

26.2.2 Robust Loss Functions

Commonly used robust alternative to LS-loss is the least absolute deviation (LAD) or L_1-loss, $\rho(e) = |e|$, which gives much larger weights on small residuals and less weight on large residuals. A compromise between LS and LAD loss function is *Huber's loss function* (Huber 1981), which is a differentiable convex combination of ℓ_2 and ℓ_1 loss functions, defined as

$$\rho_{H,c}(e) = \begin{cases} \frac{1}{2}e^2, & \text{for } |e| \le c \\ c|e| - \frac{1}{2}c^2, & \text{for } |e| > c. \end{cases} \tag{26.6}$$

Yet also Huber's loss function, as all convex loss functions, is unbounded. This implies that large outliers can still have considerable effect on the estimator. *Tukey's biweight loss function*, defined as

$$\rho_{T,c}(e) = \begin{cases} \frac{e^2}{2} - \frac{e^4}{2c^2} + \frac{e^6}{6c^4}, & \text{for } |e| \le c \\ \frac{c^2}{6}, & \text{for } |e| > c. \end{cases}$$

is bounded and continuously differentiable. Hence, it offers increased robustness against gross errors. However, it is non-convex which implies that overall minimization can get stuck in problematic local minimas. The corresponding derivative functions are $\psi_{H,c}(e) = \max[-c, \min(c, e)]$ and

$$\psi_{T,c}(e) = \begin{cases} 2e\{1 - (e/c)^2\}^2, & \text{for } |e| \le c \\ 0, & \text{for } |e| > c \end{cases}$$

and these will be needed in the Algorithm 2 (HUB-SNIHT) in the sequel. Note that $\psi_{H,c}$ is a *winsorizing (clipping) function*: the smaller value of c indicates more downweighting to the residuals, illustrating the role c on robustness. Tukey's biweight is redescending to zero which implies that large outliers can have zero weight. The above loss functions are depicted in Fig. 26.1.

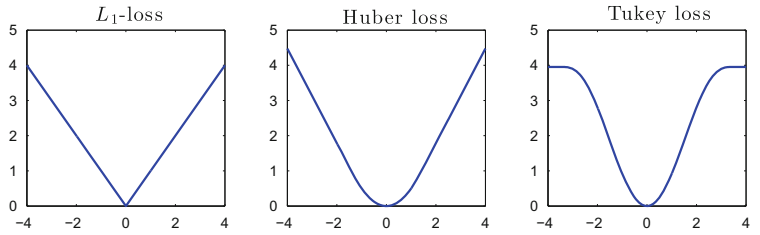

Fig. 26.1 L_1, Huber's and Tukey's ρ functions. Huber's loss function is convex but unbounded whereas Tukey's loss function is non-convex, but bounded

Note that both Huber's and Tukey's loss functions depend on a user-defined *threshold constant* c that influences the degree of robustness and efficiency of the method. The tuning parameters

$$c_{0.95} = 1.345 \quad \text{and} \quad c_{0.85} = 0.732 \tag{26.7}$$

yield 95 and 85 percent asymptotic relative efficiency of Huber's loss function compared to the LS-estimate in the multiple linear regression model with Gaussian errors; see Maronna et al. (2006).

26.2.3 Sparse Recovery Using Huber's Criterion

Note that the standard robust loss functions such as Huber's or Tukey's require a robust preliminary scale estimate $\hat{\sigma}$. Since such an estimate is difficult to obtain, we plan to estimate the signal matrix $\mathbf{X}$ and the error scale σ simultaneously. Let us first define a matrix norm as

$$\|\mathbf{X}\|_{\sigma,c} = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{Q} \rho\left(\frac{x_{ij}}{\sigma}\right)}$$

where $\sigma > 0$ is a scale parameter and ρ is assumed to be differentiable convex loss function. Due to assumption of convexity, Tukey's loss function is not permitted here. We then seek to minimize

$$\mathscr{L}(\mathbf{X}, \sigma) = \sigma \|\mathbf{Y} - \boldsymbol{\Phi}\mathbf{X}\|_{\sigma,c}^2 + (MQ)\alpha\sigma \tag{26.8}$$

$$= \sigma \sum_{i=1}^{M}\sum_{j=1}^{Q} \rho\left(\frac{y_{ij} - \boldsymbol{\phi}_{(i)}^\top \mathbf{x}_j}{\sigma}\right) + (MQ)\alpha\sigma \tag{26.9}$$

subject to $\|\mathbf{X}\|_0 \le K$,

where $\alpha > 0$ is a scaling factor chosen so that the solution $\hat{\sigma}$ is Fisher-consistent for σ when error terms are i.i.d. Gaussian, $e_{ij} \sim \mathcal{N}(0, \sigma^2)$. The objective function $\mathscr{L}$ in (26.8) was proposed for joint estimation of regression and scale of (overdetermined) multiple linear regression model by Huber (1981) and is often referred to as *"Huber's proposal 2"*. Note that $\mathscr{L}(\mathbf{X}, \sigma)$ is a convex function of $(\mathbf{X}, \sigma)$ given that ρ is convex which follows from Owen (2007). Following Huber (1981, Sect. 7.8) this fact permits developing an iterative projected gradient descent algorithm for finding an approximate solution $(\hat{\mathbf{X}}, \hat{\sigma})$ of the optimization problem (26.8). This algorithm, called the HUB-SNIHT, is described in detail in Sect. 26.3.

It is instructive to consider the optimization problem without the constraint (i.e., dropping the assumption of rowsparsity) since in this case the MMV model reduces to conventional multivariate regression problem. If it is further assumed that $M > Q$, then the model is also overdetermined and hence has (due to strict convexity of $\mathscr{L}$) a unique solution. It is easy to verify similar to the multiple linear regression problem ($Q = 1$ case studied in Owen 2007) that the negative log-likelihood function in the case that $Q > 1$ is also not convex in $(\mathbf{X}, \sigma)$, whereas Huber's criterion $\mathscr{L}(\mathbf{X}, \sigma)$ is. In addition, the minimizer $\hat{\mathbf{X}}$ of $\mathscr{L}(\mathbf{X}, \sigma)$ preserves the same theoretical robustness properties (such as bounded influence function) as the minimizer in the model where σ is a fixed known parameter. The stationary point of (26.8) can be found by setting the matrix derivative of $\mathscr{L}$ w.r.t. $\mathbf{X}$ and the derivative of $\mathscr{L}$ w.r.t. σ to zero. Simple calculations then show that the minimizer $(\hat{\mathbf{X}}, \hat{\sigma})$ of $\mathscr{L}(\mathbf{X}, \sigma)$ is a solution to the M-estimating equations:

$$\boldsymbol{\Phi}^\top \psi\left(\frac{\mathbf{R}}{\sigma}\right) = \mathbf{0} \tag{26.10}$$

$$\frac{1}{MQ} \sum_{i=1}^{M} \sum_{j=1}^{Q} \chi\left(\frac{y_{ij} - \boldsymbol{\phi}_{(i)}^\top \mathbf{x}_j}{\sigma}\right) = \alpha \tag{26.11}$$

where $\mathbf{R} = \mathbf{Y} - \boldsymbol{\Phi}\mathbf{X}$, $\psi = \rho'$ and $\chi : \mathbb{R} \to \mathbb{R}_0^+$ is defined as

$$\chi(t) = \psi(t)t - \rho(t). \tag{26.12}$$

Above the notation $\psi(\mathbf{R})$ refers to element-wise application of ψ-function to its matrix valued argument, so $[\psi(\mathbf{R})]_{ij} = \psi(r_{ij})$. Thus if ρ is convex and the MMV model is overdetermined with non-sparse $\mathbf{X}$, solving the above M-estimating equations would give the unique global minimum of (26.8).

The scaling factor α in (26.8) is chosen so that the obtained scale estimate $\hat{\sigma}$ is Fisher-consistent for the unknown scale σ when $e_{ij} \sim \mathcal{N}(0, \sigma^2)$. Due to (26.11), it is set as

$$\alpha = \mathbb{E}[\chi(e)], \quad e \sim \mathcal{N}(0, 1). \tag{26.13}$$

For many loss functions, α can be computed in closed-form. For example, for Huber's function (26.6) the χ-function in (26.12) becomes

$$\chi_{H,c}(e) = \frac{1}{2}\big[\psi_{H,c}(e)\big]^2$$

and the consistency factor α can be computed in closed-form as

$$\alpha = c^2(1 - F_G(c)) + F_G(c) - 1/2 - c f_G(c),$$

where F_G and f_G denote the c.d.f and the p.d.f. of $\mathcal{N}(0, 1)$ distribution, respectively, and c is the threshold of Huber's loss function.

26.3 SNIHT Using Huber's Criterion

Iterative hard thresholding (IHT) (Blumensath and Davies 2009), and its enhancement, the normalized IHT (NIHT) (Blumensath and Davies 2010), are popular CS signal reconstruction methods. These methods were recently extended to MMV setting in Blanchard et al. (2014). The simultaneous IHT (SIHT) and simultaneous NIHT (SNIHT) methods are described in Sect. 26.3.1 and the new robust SNIHT algorithm based on Huber's criterion $\mathscr{L}(\mathbf{X}, \sigma)$ in (26.8) is described in Sect. 26.3.2.

26.3.1 Simultaneous Normalized Iterative Hard Thresholding

Let the initial value of iteration be $\mathbf{X}^0 = \mathbf{0}$. The *simultaneous IHT (SIHT) algorithm* updates the estimate of $\mathbf{X}$ by taking steps towards the direction of the negative gradient followed by projection onto the constrained (K-rowsparse) space. It iterates

$$\mathbf{X}^{n+1} = H_K(\mathbf{X}^n + \mu^n \boldsymbol{\Phi}^\top \mathbf{R}^n),$$

where $\mathbf{R}^n = \mathbf{Y} - \boldsymbol{\Phi}\mathbf{X}^n$ denote the residual (error) matrix at nth iteration, $\mu^n > 0$ denotes a stepsize and $H_K(\cdot)$ is the hard thresholding operator. Blanchard et al. (2014) derived performance guarantees of the algorithm based on the *asymmetric restricted isometry property (RIP)* of the measurement matrix $\boldsymbol{\Phi}$ and showed that the algorithm will reach a neighbourhood of the best K-term approximation. Both in the SMV setting (Blumensath and Davies 2009) and in the MMV setting (Blanchard et al. 2014), the recovery bound of (S)NIHT is linearly dependent on the ℓ_2-norm of the noise vector/matrix, i.e., on $\|\mathbf{E}\|$. Therefore, the SNIHT method will not be robust nor guaranteed to provide accurate reconstruction in heavy-tailed noise (in such cases $\|\mathbf{E}\|$ can be very large).

Algorithm 1 SNIHT algorithm (Blanchard et al. 2014).

1: **Input** MMV matrix $\mathbf{Y}$, measurement matrix $\boldsymbol{\Phi}$, sparsity K.
2: **Output** $(\mathbf{X}^{n+1}, \Gamma^{n+1})$ estimates of K-sparse signal matrix $\mathbf{X}$ and support set Γ.
3: **Initialize** $\mathbf{X}^0 = \mathbf{0}$, $\mu^0 = 0$, $n = 0$.
4: $\Gamma^0 = \mathrm{rsupp}(H_K(\boldsymbol{\Phi}^\top \mathbf{Y}))$ ▷ Computes initial support
5: **while** halting criterion false **do**
6: $\quad \mathbf{R}^n = \mathbf{Y} - \boldsymbol{\Phi}\mathbf{X}^n$
7: $\quad \mathbf{G}^n = \boldsymbol{\Phi}^\top \mathbf{R}^n$
8: $\quad \mu^{n+1} = \left\| \boldsymbol{\Phi}_{\Gamma^n} \mathbf{G}^n_{(\Gamma^n)} \right\|^{-2} \left\| \mathbf{G}^n_{(\Gamma^n)} \right\|^2$ ▷ STEPSIZE UPDATE
9: $\quad \mathbf{X}^{n+1} = H_K(\mathbf{X}^n + \mu^{n+1}\mathbf{G}^n)$ ▷ SIGNAL UPDATE
10: $\quad \Gamma^{n+1} = \mathrm{rsupp}(\mathbf{X}^{n+1})$ ▷ SUPPORT UPDATE
11: **end while**

The pseudo-code of SNIHT algorithm is given in Algorithm 1. Note that SIHT uses fixed stepsize $\mu^n = 1$ for each iteration whereas SNIHT uses optimal stepsize update (in terms of reduction in squared approximation error) at each iteration. As its benefits, the SNIHT algorithm has a faster convergence rate and increased stability compared to SIHT. The stepsize is updated in Step 8. The rationale behind this stepsize selection is described in Sect. 26.3.2. Note that the algorithm requires an initialization procedure. The initial signal matrix is the zero matrix $\mathbf{X}^0 = \mathbf{0}$, so the initial residual (for $n = 0$) matrix is the MMV matrix $\mathbf{Y}$, i.e., $\mathbf{R}^0 = \mathbf{Y} - \boldsymbol{\Phi}\mathbf{X}^0 = \mathbf{Y}$. In Step 4 an initial proxy Γ^0 of the support set is computed. This is done by finding the support set of K largest ℓ_2-row vectors of the correlation matrix $\boldsymbol{\Phi}^\top \mathbf{Y}$. The method also requires a halting criterion which generally depends on the application. We discuss some criterions also in Sect. 26.3.2.

26.3.2 Robust SNIHT for Huber's Criterion

The SNIHT algorithm for Huber's criterion $\mathscr{L}(\mathbf{X}, \sigma)$ is given in Algorithm 2. The steps 6–14 can be divided into 3 stages described below: *scale stage* consists of Steps 6 & 7 build up the scale update σ^{n+1}, *signal stage* consists of Steps 8, 9 and 13 which build up the K-sparse signal update $\mathbf{X}^{n+1}$ and the corresponding support Γ^{n+1}, and *stepsize stage* consists of Steps 10–12 which compute the optimal stepsize update for the gradient descent move in Step 13. Naturally, one also needs to specify the differentiable convex loss function ρ. We utilize Huber's loss function $\rho_{H,c}$ in (26.6) with threshold constant c given by (26.7), but some other loss function could be used as well. Note that the selected loss function ρ then also specify the score function $\psi = \rho'$ (needed in Step 8) as well as χ-function given in (26.12) (needed in Step 7). The weight function w used in Step 11 is also defined via ρ as

$$w(e) = \frac{\psi(e)}{e}.$$

Algorithm 2 HUB-SNIHT algorithm

Input MMV matrix $\mathbf{Y}$, measurement matrix $\boldsymbol{\Phi}$, sparsity K and trimming threshold c.
Output $(\mathbf{X}^{n+1}, \sigma^{n+1}, \Gamma^{n+1})$ estimates of $\mathbf{X}$, σ and Γ.
Initialize $\mathbf{X}^0 = \mathbf{0}$, $\mu^0 = 0$, $n = 0$ and compute α in (26.13)
$\{\Gamma^0, \sigma^0\}$ =InitSupport($\mathbf{Y}, c$) ▷ Computes initial support & scale
while halting criterion false **do**
  $\mathbf{R}^n = \mathbf{Y} - \boldsymbol{\Phi}\mathbf{X}^n$
  $(\sigma^{n+1})^2 = \frac{(\sigma^n)^2}{\alpha} \frac{1}{MQ} \sum_{i=1}^{M} \sum_{j=1}^{Q} \chi\left(\frac{r_{ij}^n}{\sigma^n}\right)$ ▷ SCALE UPDATE
  $\mathbf{R}_\psi^n = \psi\left(\frac{\mathbf{R}^n}{\sigma^{n+1}}\right)\sigma^{n+1}$
  $\mathbf{G}^n = \boldsymbol{\Phi}^\top \mathbf{R}_\psi^n$
  $\mathbf{B}^n = \boldsymbol{\Phi}_{\Gamma^n} \mathbf{G}_{(\Gamma^n)}$
  $\mathbf{W}^n = w\left(\frac{\mathbf{R}^n - \mu^n \mathbf{B}^n}{\sigma^{n+1}}\right)$
  $\mu^{n+1} = \|\mathbf{B}^n\|_{\mathbf{W}^n}^{-2} \langle \mathbf{R}^n, \mathbf{B}^n \rangle_{\mathbf{W}^n}$ ▷ STEPSIZE UPDATE
  $\mathbf{X}^{n+1} = H_K(\mathbf{X}^n + \mu^{n+1}\mathbf{G}^n)$ ▷ SIGNAL UPDATE
  $\Gamma^{n+1} = \mathrm{rsupp}(\mathbf{X}^{n+1})$ ▷ SUPPORT UPDATE
end while

For example, the weight function defined by Huber's loss function is $w_{H,c}(e) = \min(1, c/|e|)$.

The algorithm can be justified heuristically as follows. First, if we consider signal matrix fixed at value $\mathbf{X}^n$, then the scale is updated using (26.11) by a fixed-point iteration

$$(\sigma^{n+1})^2 = \frac{(\sigma^n)^2}{\alpha} \frac{1}{MQ} \sum_{i=1}^{M} \sum_{j=1}^{Q} \chi\left(\frac{r_{ij}^n}{\sigma^n}\right),$$

where $\mathbf{R}^n = \mathbf{Y} - \boldsymbol{\Phi}\mathbf{X}^n$ Then, when we consider σ fixed at a value $\sigma = \sigma^{n+1}$ (the value of σ at $(n+1)$th iteration), the SNIHT update of the signal matrix becomes

$$\mathbf{X}^{n+1} = H_K\big(\mathbf{X}^n + \mu^{n+1}\boldsymbol{\Phi}^\top \mathbf{R}_\psi^n\big)$$

where μ^{n+1} is the update of the stepsize at $(n+1)$th iteration and

$$\mathbf{R}_\psi^n = \psi\left(\frac{\mathbf{R}^n}{\sigma^{n+1}}\right)\sigma^{n+1}$$

is called as *pseudo-residual*. Note that $-\nabla_{\mathbf{X}} \|\mathbf{Y} - \boldsymbol{\Phi}\mathbf{X}\|_{\sigma^{n+1},c} = (\sigma^{n+1})^{-2}\boldsymbol{\Phi}^\top \mathbf{R}_\psi^n$, where the scaling constant can be absorbed to stepsize μ^{n+1}. These two alternating steps are then iteratively updated until convergence.

`InitSupport` function in Step 4 of Algorithm 2 proceeds by computing σ^0 via solving (26.14) and then computing Γ^0 using (26.15). Let us briefly discuss the

rationale behind these computations. Note that the algorithm starts with $\mathbf{X}^0 = \mathbf{0}$, so $\mathbf{R}^0 = (r_{ij}^0) = \mathbf{Y}$. Under the assumption that $\mathbf{X} = \mathbf{0}$ and that the error terms are i.i.d. with common scale σ, one may compute the simultaneous location estimate $\hat{\mu}$ and scale estimate σ^0 of the data set $\text{vec}(\mathbf{Y}) \in \mathbb{R}^{(MQ)\times 1}$ using Huber's criterion as the unique minimizer of

$$(\hat{\mu}, \sigma^0) = \underset{\mu\in\mathbb{R},\sigma>0}{\arg\min}\sigma \sum_{i=1}^{M}\sum_{j=1}^{Q} \rho\left(\frac{y_{ij}-\mu}{\sigma}\right) + (MQ)\alpha\sigma \tag{26.14}$$

which can be solved using a simple iterative algorithm given in Huber (1981, Sect. 7.8). Given the pair $(\hat{\mu}, \sigma^0)$ above, the initial support estimate Γ^0 is computed as

$$\Gamma^0 = \text{rsupp}\left(H_K(\boldsymbol{\Phi}^\top \mathbf{R}_\psi^0)\right), \tag{26.15}$$

where $\mathbf{R}_\psi^0$ is a matrix of pseudo-residuals whose (i,j)th element is given by

$$\left[\mathbf{R}_\psi^0\right]_{i,j} = \left(\psi\left(\frac{r_{ij}^0 - \hat{\mu}}{\sigma^0}\right)\sigma^0\right).$$

Alternatively, instead of solving (26.14), one can compute $\hat{\mu}$ and σ^0 as the median and median absolute deviation scale estimate of y_{ij}'s, respectively.

Let us next turn our attention to the stepsize stage. First note that, assuming we have identified the correct support at nth iteration, then we can choose the stepsize update μ^{n+1} as the minimizer of $\|\mathbf{Y} - \boldsymbol{\Phi}\mathbf{X}\|_{\sigma,c}^2$ for fixed scale at $\sigma = \sigma^{n+1}$ in the gradient ascent direction $\mathbf{X}^n + \mu\mathbf{G}^n|_{\Gamma^n}$. Thus we find μ as the minimizer of

$$L(\mu) = \|\mathbf{Y} - \boldsymbol{\Phi}\,(\mathbf{X}^n + \mu\mathbf{G}^n|_{\Gamma^n})\|_{\sigma^{n+1},c}^2 = \|\mathbf{R}^n - \mu\mathbf{B}^n\|_{\sigma^{n+1},c}^2$$

where $\mathbf{B} = \boldsymbol{\Phi}_{\Gamma^n}\mathbf{G}^n_{(\Gamma^n)}$ and $\mathbf{R}^n = \mathbf{Y} - \boldsymbol{\Phi}\mathbf{X}^n$. This is an M-estimation problem of a simple (one predictor) linear regression model with response vector $\mathbf{r} = \text{vec}(\mathbf{R}^n)$ and regressor $\mathbf{b} = \text{vec}(\mathbf{B}^n)$. A standard approach for finding the (unique) minimizer of $L(\mu)$ is the *iteratively reweighted LS (IRWLS) algorithm* which iterates the steps

$$\mathbf{W} \leftarrow w\left(\frac{\mathbf{R}^n - \mu\mathbf{B}^n,}{\sigma^{n+1}}\right), \quad \mu \leftarrow \|\mathbf{B}^n\|_{\mathbf{W}}^{-2}\,\langle \mathbf{R}^n, \mathbf{B}^n\rangle_{\mathbf{W}} \tag{26.16}$$

until convergence given some initial value of μ to start the iteration. (Note that we have expressed the IRWLS steps using more convenient matrix notations rather than vector notations via $\mathbf{r}$ and $\mathbf{b}$). Then, instead of choosing the next update μ^{n+1} as the minimizer of $L(\mu)$ by iterating the steps in (26.16) until convergence, we use a 1-step estimator which corresponds to a single iteration of the steps above with initial value of iteration given by the previous stepsize μ^n. Hence, the steps 10–12

(stepsize stage) of Algorithm 2 form a single run of IRWLS steps in (26.16). In our simulations we noticed that 1-step iterate already gives a very good estimate, often within 3 decimal accuracy, and hence further iterations were not required.

Finally, the algorithm also requires a halting criterion. Some possibilities are:

1. Halt the algorithm when the maximum total (squared) correlation between an atom and the pseudo-residual falls below a level δ: $\|\boldsymbol{\Phi}^\top \mathbf{R}_\psi^n\|_{2,\infty} < \delta$.
2. Stop the algorithm when previous and current iterations are close to each other in the sense that $\frac{\|\mathbf{X}^{n+1}-\mathbf{X}^n\|^2}{\|\mathbf{X}^{n+1}\|^2} < \delta$ for a given level δ

In our simulation study, we used the stopping criterion 2 with $\delta = 1.0^{-8}$.

26.4 Simulation Studies

Next we provide a set of simulation studies to illustrate the usefulness of the proposed HUB-SNIHT method in a variety of noise environments and SNR levels. Also, the effect of number of measurement vectors Q on exact support recovery probability will be illustrated.

26.4.1 Set-Up

The elements of $\boldsymbol{\Phi}$ are drawn from $\mathcal{N}(0,1)$ distribution and the columns are normalized to have a unit norm. The coefficients of K nonzero row vectors of $\mathbf{X}$ have equal amplitudes $\sigma_x = |x_{ij}| = 10 \forall i \in \Gamma, j = 1,\ldots,Q$ and equiprobable signs (i.e., $\pm 1 = \text{sign}(x_{ij})$ with equal probability $\frac{1}{2}$) and the signal support set $\Gamma = \text{supp}(\mathbf{X})$ is randomly chosen from $\{1,\ldots,N\}$ without replacement for each trial. In all of our experiments, the noise matrix $\mathbf{E}$ consists of i.i.d. elements e_{ij} from a continuous symmetric distribution F_e with scale parameter σ. We define the (generalized) *signal to noise ratio (SNR)* as $\text{SNR}(\sigma) = 20\log_{10}\frac{\sigma_x}{\sigma}$ which depends on the used scale parameter σ of the error distribution.

As performance measures of sparse signal recovery, we use both the (observed) *mean squared error*

$$\text{MSE}(\hat{\mathbf{X}}) = \frac{1}{LQ}\sum_{\ell=1}^{L}\left\|\hat{\mathbf{X}}^{[\ell]} - \mathbf{X}^{[\ell]}\right\|^2$$

and the empirical *probability of exact recovery (PER)*

$$\text{PER} \triangleq \frac{1}{L}\sum_{\ell=1}^{L}\text{I}\left(\hat{\Gamma}^{[\ell]} = \Gamma^{[\ell]}\right)$$

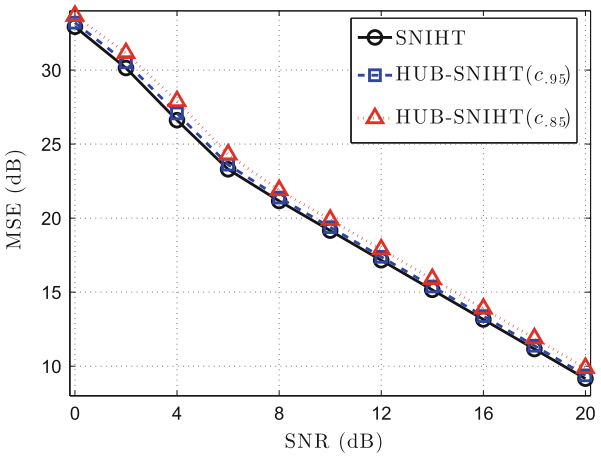

Fig. 26.2 Average MSE of SNIHT methods as a function of SNR in $\mathcal{N}(0, \sigma^2)$ noise; System parameters were $(M, N, K, Q) = (256, 512, 8, 16)$ and the number of trials is 4000

where $\mathrm{I}(\cdot)$ denotes the indicator function, $\hat{\mathbf{X}}^{[\ell]}$ and $\hat{\Gamma}^{[\ell]} = \mathrm{rsupp}(\hat{\mathbf{X}}^{[\ell]})$ denote the estimate of the K-sparse signal $\mathbf{X}^{[\ell]}$ and the signal support $\Gamma^{[\ell]}$ for the ℓth Monte Carlo (MC) trial, respectively. In all simulation settings described below, the number of MC trials is $L = 4000$, the length of the signal is $N = 512$, the number of measurements is $M = 256$, and the sparsity level is $K = 8$. The number of measurement vectors is $Q = 16$ unless otherwise stated. The methods included in they study are **SNIHT** described in Algorithm 1 and **HUB-SNIHT** of Algorithm 2 using the Huber's loss function $\rho_{H,c}(\cdot)$ with trimming threshold $c_{0.95}$ or $c_{0.85}$ given in (26.7).

26.4.2 Results

Gaussian Noise Case Figure 26.2 shows the signal reconstruction performance in terms of MSE as a function of SNR in i.i.d. Gaussian noise, $e_{ij} \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = \mathbb{E}[e_{ij}^2]$. As expected, the SNIHT has the best performance, but HUB-SNIHT with $c_{0.95}$ (resp. $c_{0.85}$) suffers a negligible 0.2 dB (resp. 0.65 dB) performance loss. Both SNIHT and HUB-SNIHT methods had a full PER rate ($= 1$) for the SNR levels ≥ 6 dB. The PER rate declines for < 6 dB and reaches 0 at SNR $= 0$ dB for all of the methods. The usefulness of joint recovery becomes more evident at low SNR's, where multiple measurements can dramatically improve on the recovery ability by exploiting the joint information. This is illustrated in our 2nd simulation experiment, where SNR is fixed at low 6 dB level and we let the number of measurements Q vary $Q \in \{2, 4, \ldots, 24\}$. Figure 26.3 depicts the empirical PER rates as a function of number of measurements. Again the number of trials is 4000 for each parameter set.

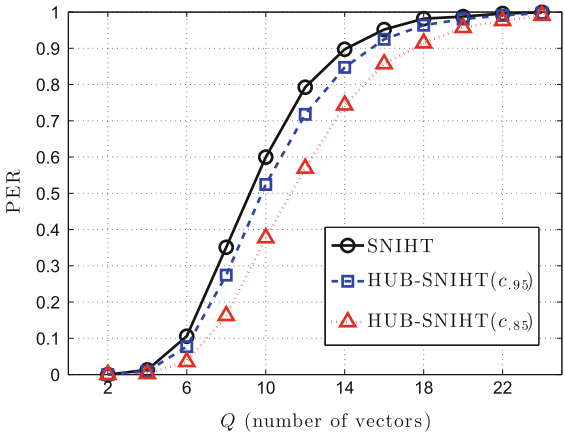

Fig. 26.3 Empirical PER rates of SNIHT methods as a function of number of measurements Q in $\mathcal{N}(0, \sigma^2)$ noise; $(M, N, K) = (256, 512, 8)$, SNR $= 6$ dB, and the number of trials is 4000

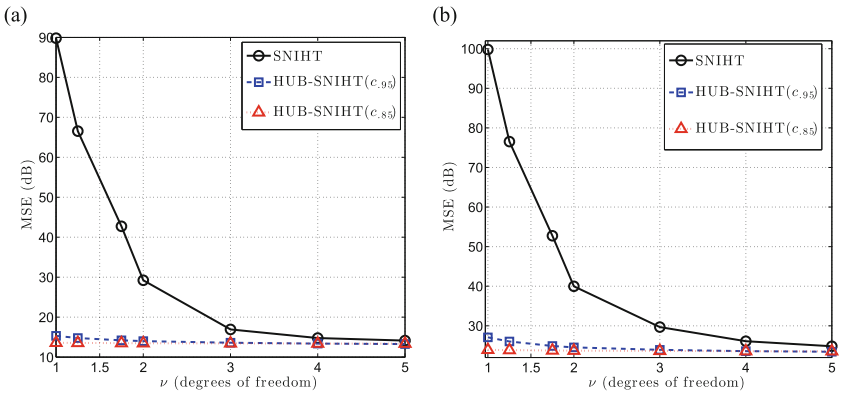

Fig. 26.4 Average MSE of SNIHT methods in $t_\nu(0, \sigma)$ distributed noise as a function of d.o.f. ν and different SNR levels; $(M, N, K, Q) = (256, 512, 8, 16)$ and the number of trials is 4000. (**a**) SNR$(\sigma) = 20$ dB. (**b**) SNR$(\sigma) = 10$ dB

Note that the PER rate increases as a function of Q from total failure 0 (when $Q = 2$) to near full 100 % recovery when $Q = 22$. HUB-SNIHT using $c_{0.95}$ suffers only a small performance loss compared to SNIHT, but HUB-SNIHT based on threshold $c_{0.85}$ is considerably behind the other two methods.

Student's t_ν-Distributed Noise Figure 26.4 depicts the MSE in i.i.d. t-distributed noise, $e_{ij} \sim t_\nu(0, \sigma^2)$ as a function of degrees of freedom (d.o.f.) ν. The scale parameter is $\sigma = \mathrm{Med}_{F_e}(|e_{ij}|)$ and SNR(σ) was 10 and 20 dB. Recall that t-distribution with $\nu > 0$ degrees of freedom (d.o.f.) is heavy-tailed distribution with $\nu = 1$ corresponding to Cauchy distribution and $\nu \to \infty$ corresponding to Gaussian distribution. The proposed HUB-SNIHT method has solid performance for both

Table 26.1 PER rates in $t_\nu(0,\sigma)$ noise at varying SNR(σ) and d.o.f. ν

		Degrees of freedom ν							
SNR(σ)	Method	1	1.25	1.5	1.75	2	3	4	5
20 dB	SNIHT	0	0	0.07	0.43	0.78	1.0	1.0	1.0
	HUB-SNIHT $c_{0.95}$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	HUB-SNIHT $c_{0.85}$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
10 dB	SNIHT	0	0	0	0	0	0.09	0.50	0.74
	HUB-SNIHT $c_{0.95}$	0.32	0.50	0.65	0.73	0.80	0.88	0.93	0.94
	HUB-SNIHT $c_{0.85}$	0.88	0.90	0.91	0.91	0.92	0.92	0.93	0.93

System parameters were $(M,N,K,Q) = (256,512,8,16)$ and the results are averages over 1000 trials. For SNR 20 dB, the proposed HUB-SNIHT methods are able to maintain full recovery rates even in Cauchy noise ($\nu = 1$). At low SNR 10 dB, the conventional SNIHT breaks down completely

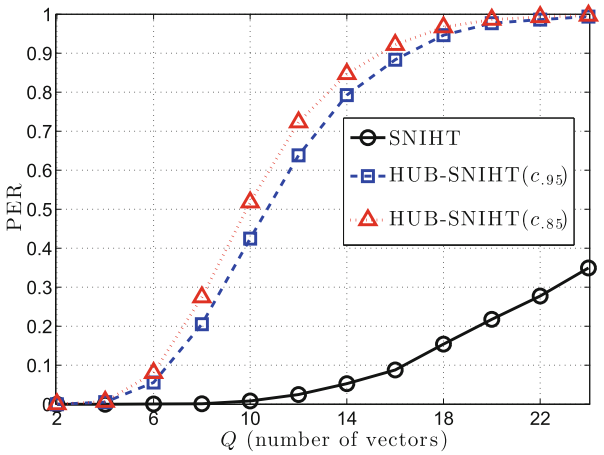

Fig. 26.5 Empirical PER rates of SNIHT methods as a function of number of measurements Q in $t_3(0,\sigma^2)$ noise (i.e., $\nu = 3$); $(M,N,K) = (256,512,8)$, SNR(σ) $= 10$ dB, the number of trials is 4000

SNR levels whereas SNIHT has an exponential increase in MSE as ν decreases. At high SNR 20 dB, the HUB-SNIHT methods are able to retain a steady MSE for all values of ν. Overall, the HUB-SNIHT method using $c_{0.95}$ is (as expected) less robust with slightly worse performance. The PER rates in Table 26.1 illustrate the remarkable performance of HUB-SNIHT methods. At high SNR $= 20$ dB the robust methods are able to maintain full 100 % recovery rates even when the noise has impulsive Cauchy distribution ($\nu = 1$). At low SNR 10 dB, the HUB-SNIHT with $c_{0.85}$ has the best performance as expected and is able to maintain good PER rates (always above 88 %) whereas SNIHT fails completely for $\nu \leq 3$. In our 4th simulation experiment, we again illustrate the usefulness of joint recovery. SNR is fixed at 10 dB, the noise follows t-distribution with $\nu = 3$ d.o.f. and number of measurements Q vary $Q \in \{2,4,\ldots,24\}$. Figure 26.5 shows the empirical PER

rates as a function of number of measurements. In the case of HUB-SNIHT, the PER rates increase as a function of Q from total failure 0 (when $Q = 2$) to full 100 % recovery (when $Q = 24$). The PER rates of SNIHT are drastically behind the robust methods, reaching the highest PER rate of 36 % at $Q = 24$. This is in deep contrast with 100 % recovery obtained by HUB-SNIHT methods.

26.5 Conclusions

In this chapter, we developed a robust simultaneous iterative hard thresholding (SNIHT) algorithm for the MMV model, also known as multichannel sparse recovery problem. Our approach is based on generalizing the Huber's criterion for joint estimation of location and scale to MMV model. A set of simulations studies illustrated that the method gives very similar performance in Gaussian errors as conventional SNIHT, but superior performance in heavy-tailed noise.

Acknowledgements The author wishes to thank Academy of Finland for supporting this research.

References

Blanchard, J.D., Cermak, M., Hanle, D., Jin, Y.: Greedy algorithms for joint sparse recovery. IEEE Trans. Signal Process. **62**(7), 1694–1704 (2014)

Blumensath, T., Davies, M.E.: Iterative hard thresholding for compressed sensing. Appl. Comput. Harmon. Anal. **27**(3), 265–274 (2009)

Blumensath, T., Davies, M.E.: Normalized iterative hard thresholding: guaranteed stability and performance. IEEE J. Sel. Top. Sign. Proces. **4**(2), 298–309 (2010)

Candes, E.J., Wakin, M.B.: An introduction to compressive sampling. IEEE Signal Process. Mag. **25**(2), 21–30 (2008)

Carrillo, R., Barner, K.: Lorentzian based iterative hard thresholding for compressed sensing. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'11), pp. 3664–3667. Prague, Czech Republic (2011)

Chen, J., Huo, X.: Theoretical results on sparse representations of multiple-measurement vectors. IEEE Trans. Signal Process. **54**(12), 4634–4643 (2006)

Cotter, S., Rao, B.: Sparse channel estimation via matching pursuit with application to equalization. IEEE Trans. Commun. **50**(3), 374–377 (2002)

Donoho, D.: Compressive sensing. IEEE Trans. Inf. Theory **52**(2), 5406–5425 (2006)

Duarte, M.F., Eldar, Y.C.: Structured compressed sensing: from theory to applications. IEEE Trans. Signal Process. **59**(9), 4053–4085 (2011)

Elad, M.: Sparse and Redundant Representations. Springer, New York (2010)

Eldar, Y.C., Rauhut, H. Average case analysis of multichannel sparse recovery using convex relaxation. IEEE Trans. Inf. Theory **56**(1), 505–519 (2010)

Gorodnitsky, I., George, J., Rao, B. Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm. J Electroencephalogr. Clin. Neurophysiol. **95**(4), 231–251 (1995)

Gribonval, R., Zibulevsky, M. Handbook of Blind Source Separation, pp. 367–420. Academic, Oxford (2010)

Gribonval, R., Rauhut, H., Schnass, K., Vandergheynst, P.: Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms. J. Fourier Anal. Appl. **14**(5–6), 655–687 (2008)

Huber, P.J.: Robust Statistics. Wiley, New York (1981)

Malioutov, D., Çetin, M., Willsky, A.S.: A sparse signal reconstruction perspective for source localization with sensor arrays. IEEE Trans. Signal Process. **53**(8), 3010–3022 (2005)

Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust Statistics: Theory and Methods. Wiley, New York (2006)

Ollila, E.: Nonparametric simultaneous sparse recovery: an application to source localization. In: Proceedings of the European Signal Processing Conference (EUSIPCO'15). Nice, France (2015). http://arxiv.org/abs/1502.02441

Ollila, E., Kim, H.J., Koivunen, V.: Robust compressive sensing. In: Regularization, Optimization, Kernels, and Support Vector Machines. Machine Learning and Pattern Recognition Series, pp 217–236. Chapman and Hall/CRC. Boca Raton, USA (2014a)

Ollila, E., Kim, H.J., Koivunen, V. Robust iterative hard thresholding for compressed sensing. In: Proceedings of the IEEE International Symposium on Communications, Control and Signal Processing (ISCCSP'14), pp. 226–229. Athens, Greece (2014b)

Ou, W., Hämäläinen, M.S., Golland, P. A distributed spatio-temporal EEG/MEG inverse solver. Neuroimage **44**(3), 932–946 (2009)

Owen, A.B.: A robust hybrid of lasso and ridge regression. Contemp. Math. **443**, 59–72 (2007)

Razavi, S.A., Ollila, E., Koivunen, V.: Robust greedy algorithms for compressed sensing. In: Proceedings of the 20th European Signal Processing Conference (EUSIPCO'12), pp 969–973. Bucharest, Romania (2009)

Tibshirani, R. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**, 267–288 (1996)

Tropp, J.A.: Algorithms for simultaneous sparse approximation. part II: convex relaxation. Signal Process. **86**, 589–602 (2006)

Tropp, J.A., Gilbert, A.C., Strauss, M.J.: Algorithms for simultaneous sparse approximation. part I: greedy pursuit. Signal Process. **86**, 572–588 (2006)

Turlach, B., Venables, W.N., Wright, S.J.: Simultaneous variable selection. Technometrics **47**(3), 349–363 (2005)

Chapter 27
Nonparametric Detection of Complex-Valued Cyclostationary Signals

Visa Koivunen and Jarmo Lundén

Abstract In wireless communication, radar, surveillance, and active sensing systems it is necessary to detect transmitted signals in noise. Employed signals are man-made and consequently have many statistical and structural properties that can be used to aid the detection. Such properties are present even if the transmitted data itself would be random or unknown. One such key property is cyclostationarity which means that some signal statistics such as autocorrelation function are periodic. Typically these signals are observed in demanding signal environments where the standard assumption on Additive White Gaussian Noise (AWGN) may not be true. Signal detection has to be performed reliably in the face of interference and in challenging propagation environments characterized by shadowing and fading effects as well as heavy-tailed noise distributions. In this chapter, a robust computationally efficient nonparametric detector of second order cyclostationary statistics based on the spatial sign function is presented. Nonparametric approach leads to improved robustness against heavy-tailed noise and in cases when the noise statistics are not fully specified. Asymptotic distribution of the spatial sign cyclic correlation estimator under the null hypothesis is established. Tests using constraint on false alarm rate constraint are derived for single detector and decentralized detector employing multiple distributed sensors. A theoretical justification for spatial sign based detection of cyclostationary signals is provided. A sequential test for reducing the average detection time and detecting rapid changes is presented. Simulation examples on identifying idle radio spectrum are provided. Simulation example shows the reliable and statistically robust performance of the proposed nonparametric detector both in heavy-tailed and Gaussian noise environments.

Keywords Cognitive radio • Cyclostationary • Man-made signals • Nonparametric test • Robust statistics • Spectrum sensing

V. Koivunen (✉) • J. Lundén
Department of Signal Processing and Acoustics, Aalto University, 00076 Aalto, Finland
e-mail: visa.koivunen@aalto.fi; jarmo.lunden@aalto.fi

K. Nordhausen, S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, DOI 10.1007/978-3-319-22404-6_27

27.1 Introduction

Many hypothesis testing methods employed for detecting man-made signals exploit the cyclostationarity property of the signals. Most man-made signals of interest signals are complex-valued and convey information about both the phase and amplitude of the signal. Cyclostationarity in wireless communication or radar signals may be caused by sampling, multiplexers, modulation or coding and may occur, for example, at the chip or symbol rate or at the doubled carrier frequency. Cyclostationary statistics are not phase blind unlike power spectrum. Hence, they can be used for blind channel equalization and characterizing the phase component of a waveform as well. We present a statistically robust method for detecting cyclostationary wireless communication signals in noise. The problem is modeled as a binary hypothesis testing problem. Cyclostationarity-based detection allows automatic classification of signals exhibiting cyclostationarity property at different cyclic frequencies. This may be used in signal intelligence, automatic classification of communication and radar signals as well as spectrum surveillance, for example. Moreover, in cognitive radio systems it allows for distinguishing among primary licensed users, secondary users, and interference, thus preparing the way for reliable and efficient co-existence of the primary and secondary users. Moreover, random noise is commonly not cyclostationarity which makes cyclostationary signal analysis less sensitive to noise.

Wide sense Cyclostationary processes are stochastic processes whose statistical properties such as the mean and autocorrelation functions are periodic in time, i.e.

$$m_x(t) = m_x(t+T) \tag{27.1}$$

$$R_x(t_1, t_2) = R_x(t_1 + T, t_2 + T) \tag{27.2}$$

where T is the period. We are mainly interested in the autocorrelation function in this study. Because it is periodic it may be represented using Fourier series

$$R_x(t + \frac{\tau}{2}, t - \frac{\tau}{2}) = \sum_{\alpha} R_x^{\alpha}(\tau) e^{j2\pi\alpha t} \tag{27.3}$$

where $R_x^{\alpha}(\tau)$ are the Fourier coefficients

$$R_x^{\alpha}(\tau) = \frac{1}{T} \int_{-T/2}^{T/2} R_x(t + \frac{\tau}{2}, t - \frac{\tau}{2}) e^{-j2\pi\alpha t} dt \tag{27.4}$$

and α is called cyclic frequency. A cyclic autocorrelation function is depicted in Fig. 27.1 with cyclic frequencies $\alpha = \pm 0.4$. The cyclic autocorrelation function is two-dimensional with parameters τ and α.

Statistical robustness of detection algorithms is motivated by several measurement studies that demonstrate the impulsive nature of man-made noise in radio frequency bands outdoors and indoors, see Middleton (1999). Typical noise

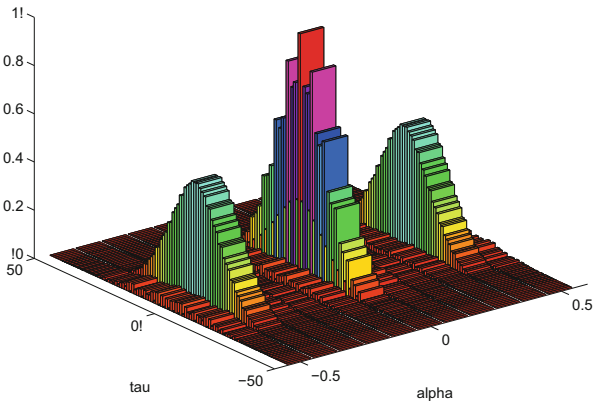

Fig. 27.1 Cyclostationary signals: cyclic autocorrelation function is not identically zero for some cycle frequency $\alpha \neq 0$. In the figure $\alpha = \pm 0.4$

distribution observed in these studies has significantly heavier tails than Gaussian distribution. Impulsive noise or interference may be due to the presence of multiple interferers contaminating the signal of interest we wish to detect. For example, indoor measurements campaigns in the ISM bands clearly show the impulsive nature of the noise and interference due to, e.g., microwave ovens and devices with electromechanical switches (including copy machines, electric motors in elevators, printers, etc.). Moreover, on a computer board various components, such as the LCD pixel clock and the PCI express bus, induce impulsive interference that degrades the performance of the embedded wireless devices. The impulsive interference observed on a computer platform is well modeled by a symmetric alpha-stable distribution. As an example of outdoor scenarios in urban environments, impulsive noise measurements on a digital television band have been reported in Sánchez et al. (1999). Additional man-made impulsive noise sources in urban outdoor radio environments include igniting car engines, power lines, and heavy current switches.

In this chapter, a nonparametric cyclic correlation estimator based on the multivariate (spatial) sign function in Visuri at al. (2000) and Kassam (1993) applied to complex-valued signals is presented. We demonstrate that the cyclostationarity property employed in the hypothesis testing is preserved under the spatial sign function for an important special case, circularly symmetric complex normal distribution. Most of the results and methods in this chapter including asymptotic distribution of the estimator under the null hypothesis have been established in Lunden et al. (2010). Test statistics for an individual and decentralized detectors are provided. A sequential method for rapid change detection and reducing the average detection time is presented as well. Simulation experiments in finding idle radio spectrum illustrate the robust performance of the detectors in non-Gaussian noise environments as well as good performance in Gaussian noise environments is presented. The methods considered are based on nonparametric statistics making them highly attractive in real-world applications where noise

and interference statistics may not be explicitly known or may change frequently (Kassam 1988, Kassam 1993). No additional nuisance parameters such as the scale need to be estimated unlike in the M-estimation based robust methods. Furthermore, nonparametric detectors achieve a fixed false alarm rate under all conditions satisfying the nonparametric null hypothesis. The detectors based on spatial sign function are computationally very efficient which makes them very attractive for hardware implementation in real-world radio frequency sensors.

27.2 Spatial Sign Cyclic Correlation

The multivariate generalization of sign function, i.e. the spatial sign for complex-valued data $x(t)$, is defined as (Visuri at al. 2000)

$$S(x(t)) = \begin{cases} \frac{x(t)}{|x(t)|}, & x(t) \neq 0 \\ 0, & x(t) = 0, \end{cases} \tag{27.5}$$

where $|\cdot|$ stands for the modulus of the complex-valued argument.

In order to study the cyclostationarity of a signal, we define the spatial sign cyclic correlation estimator as

$$\hat{R}_S(\alpha, \tau) = \frac{1}{M} \sum_{t=0}^{M-1} S(x(t)) S(x^*(t+\tau)) e^{-j2\pi\alpha t}, \ \forall \tau \neq 0 \tag{27.6}$$

where $x(t)$ is a discrete-time signal (sequence), $x^*(t)$ its complex conjugate, τ is a discrete time delay, M is the number of data samples, and α denotes the cyclic frequency.

In (27.6) it has been assumed that the signal is centered, so that the mean (or median) of the data is zero. Most radio transceivers produce such data. Otherwise, location parameters would have to be estimated and subtracted from the received signal before applying the estimator.

The use of the spatial sign function leads to an estimator that is both qualitatively and quantitatively robust, i.e. the influence of outlying observations is bounded and the method has a high breakdown point, i.e. tolerates a high proportion of contaminated data. Qualitative robustness was verified by establishing the influence function (IF). It was shown to be uniformly bounded for the spatial sign covariance estimator (Croux et al. 2002). This result can be easily extended to spatial sign based cyclic covariances at non-zero cyclic frequencies since a non-zero cyclic frequency adds only an exponential multiplier term $e^{-j2\pi\alpha t}$ to the covariance estimator that has no impact on the magnitude of the observation pairs. Consequently, the influence function possesses the uniformly bounded property.

It is important to understand what is the effect of the spatial sign function to the cyclic frequencies of the signal of interest. In practice, this may have to be

determined by either using analytical tools or experimentally for each different class of signals of interest. It has been shown that the periodicity and cyclic frequencies of the autocorrelation function are preserved for a circularly symmetric complex Gaussian process when the spatial sign function is employed. As an important special case, a multicarrier OFDM (orthogonal frequency division multiplexing) signal is approximately normal distributed particularly when the number of subcarriers is large, as in the case of digital TV signals (DVB-T, DVB-T2). We will show that the spatial sign function preserves the cyclic frequencies of a CP (cyclic prefix) OFDM system with a simulation example. OFDM is modulation method of choice in many existing and future wireless communications systems, such as the IEEE 802.11a/g WLANs, terrestrial digital television DVB-T, and 4G LTE, among others.

Detection of cyclostationary wireless communication signals is formulated as a binary hypothesis testing problem. In order to define a hypothesis test that satisfies a constraint on the false alarm rate for the presence of cyclostationarity at a given cyclic frequency, the distribution of the estimator needs to be established. In the next section, the distribution of the spatial sign cyclic correlation estimator is derived for independent and identically distributed (i.i.d.) complex circular noise process with zero mean. Circularity describes the statistical relations of a complex random variable and its complex conjugate, see Ollila et al. (2012). Nonparametric performance is achieved for all i.i.d. circular noise pdfs with zero mean. As already noted if the mean is not zero it can be estimated and subtracted from the data. However, wireless transceivers practically always produce such centered data. Note that circularity is not required from the primary user wireless signal that we wish to detect.

27.3 Statistical Properties of the Spatial Sign Cyclic Correlation Estimator

In typical sensing applications the number of observations M is typically large (in the order of several thousands). Proper Nyquist rate sampling for a millisecond would result in several thousands of observations. Hence, we may apply a particular central limit theorem to infer the distribution of the spatial sign cyclic correlation estimator.

Central limit theorem for m-dependent variables states that the sum of m-dependent random variables with finite third absolute moment has a limiting Gaussian distribution as the number of random variables goes to infinity (Fraser 1957). A sequence of random variables is defined to be m-dependent if any subset $(X_1, X_2, \ldots, X_r)$ is independent of any other subset $(X_s, X_{s+1}, \ldots)$ provided that $s - r > m$. A sequence of spatial sign correlation samples $S(x(t))S(x^*(t+\tau))$ formed from i.i.d. samples are m-dependent with $m = \tau$. Therefore, according to the central limit theorem the distribution of the spatial sign cyclic correlation estimator approaches Gaussian distribution as M goes to ∞. Thus, we can employ

a Gaussian approximation for large M and the asymptotic distribution is fully specified by determining its mean and covariance. Validity of the central limit theorem approximation will be verified by simulations in Sect. 27.5.

We assume that $x(t) = n(t)$ where $n(t)$ is a circular independent and identically distributed (i.i.d.) noise process with zero mean. That is, only noise is considered to be present. In that case, the mean of $\hat{R}_S(\alpha, \tau)$ is given by ($\tau \neq 0$)

$$E[\hat{R}_S(\alpha,\tau)] = \frac{1}{M}\sum_{t=0}^{M-1} E[S(n(t))S(n^*(t+\tau))]e^{-j2\pi\alpha t} \tag{27.7}$$

$$= 0, \ \forall\alpha \tag{27.8}$$

where the noise is assumed to be independent and circular. Consequently, $S(n(t)) = e^{j\phi}$ where ϕ has a uniform distribution between 0 and 2π.

For the zero mean case, the covariance of $\hat{R}_S(\alpha, \tau)$ and $\hat{R}_S(\beta, \rho)$ is given by ($\tau \neq 0, \rho \neq 0$) (Lunden et al. 2010)

$$\begin{aligned}
&\mathrm{Cov}(\hat{R}_S(\alpha,\tau),\hat{R}_S^*(\beta,\rho)) = E[(\hat{R}_S(\alpha,\tau))(\hat{R}_S(\beta,\rho))^*] \\
&= E\left[\left(\frac{1}{M}\sum_{t=0}^{M-1} S(n(t))S(n^*(t+\tau))e^{-j2\pi\alpha t}\right)\right. \\
&\qquad \left.\cdot\left(\frac{1}{M}\sum_{k=0}^{M-1} S(n(k))S(n^*(k+\rho))e^{-j2\pi\beta k}\right)^*\right] \\
&= \frac{1}{M^2}\sum_{t=0}^{M-1}\sum_{k=0}^{M-1} E[S(n(t))S(n^*(k)) \\
&\qquad \cdot S(n^*(t+\tau))S(n(k+\rho)))]e^{-j2\pi(\alpha t-\beta k)} \\
&= \frac{1}{M^2}\left(\sum_{t=0}^{M-1} E[|S(n(t))|^2 S(n^*(t+\tau))S(n(t+\rho))]\right. \\
&\quad \cdot e^{-j2\pi(\alpha-\beta)t} + \sum_{t=0}^{M-1}\sum_{\substack{k=0\\k\neq t}}^{M-1} E[S(n(t))S(n^*(k)) \\
&\quad \left.\cdot S(n^*(t+\tau))S(n(k+\rho))]e^{-j2\pi(\alpha t-\beta k)}\right).
\end{aligned} \tag{27.9}$$

We will consider the two sums above separately. For the i.i.d. and circular noise, the first sum is zero if $\tau \neq \rho$. When $\tau = \rho$, the expectation $E[|S(n(t))|^2|S(n(t+\tau))|^2] = 1$. In case of the second sum, the noise are i.i.d. and circular, the expectation can be non-zero only if $t = k + \rho$ and $k = t + \tau$, that is, when $\tau = -\rho$. However in

that case, the expectation may be written as

$$E[S(n(t))S(n^*(t+\tau))S(n^*(t+\tau))S(n(t))]$$

which is zero as well since $n(t)$ is circular, and thus $E[S(n(t))S(n(t))] = 0, \forall t$. Consequently, the covariance of $\hat{R}_S(\alpha, \tau)$ for a circular i.i.d. process with zero mean may be written as follows for $\tau \neq 0, \rho \neq 0$:

$$\begin{aligned} &\mathrm{Cov}(\hat{R}_S(\alpha,\tau), \hat{R}_S^*(\beta,\rho)) \\ &= \begin{cases} \frac{1}{M} & ,\alpha = \beta, \tau = \rho \neq 0, \\ \frac{1}{M^2} \sum_{t=0}^{M-1} e^{-j2\pi(\alpha-\beta)t} & ,\alpha \neq \beta, \tau = \rho \neq 0, \\ 0 & ,\tau \neq \rho. \end{cases} \end{aligned} \tag{27.10}$$

27.4 Non-parametric Test for Cyclostationary Signals

We formulate the null and alternative hypotheses as follows:

$$\begin{aligned} H_0 &: x(t) = n(t), \\ H_1 &: x(t) = s(t) + n(t), \end{aligned} \tag{27.11}$$

where $x(t)$ is the received signal, $s(t)$ is the transmitted man-made signal of interest that has possibly passed through a propagation medium with potentially time-varying impulse response, and $n(t)$ denotes complex, circular i.i.d. noise.

Under the null hypothesis H_0, the expected value of the estimated spatial sign cyclic correlation is zero and noise only is observed. Under the alternative hypothesis H_1, the expected value of the estimated spatial sign cyclic correlation is non-zero for the cyclic frequencies of the complex-valued signal of interest that is processed using spatial sign function.

Let $\hat{\boldsymbol{r}}_S(\alpha)$ denote a vector that contains the estimated spatial sign cyclic correlations at cyclic frequency α for a set of time delays $\tau_1, \ldots, \tau_N$,

$$\hat{\boldsymbol{r}}_S(\alpha) = [\hat{R}_S(\alpha, \tau_1), \ldots, \hat{R}_S(\alpha, \tau_N)]^T. \tag{27.12}$$

From Sect. 27.3, it follows that under the null hypothesis for large M, estimated cyclic correlations obey the distribution

$$\hat{\boldsymbol{r}}_S(\alpha) \sim N_C(\boldsymbol{0}, \frac{1}{M}\boldsymbol{I}), \ \forall\alpha, \ \forall\tau_i \neq 0, i = 1, \ldots, N, \tag{27.13}$$

where $N_C(\cdot,\cdot)$ denotes the complex Gaussian distribution and $\boldsymbol{I}$ the identity matrix, respectively.

Now, we define the test statistic for the spatial sign cyclic correlation based test for a single sensor (one individual detector) as

$$\lambda = M||\hat{\boldsymbol{r}}_S(\alpha)||^2, \tag{27.14}$$

where $||\cdot||$ denotes the Euclidean norm. Since the $\hat{R}_S(\alpha, \tau)$ are complex Gaussian distributed, the test statistic λ is the log-likelihood under the null hypothesis after neglecting the constant terms. The null hypothesis is rejected if $\lambda > \gamma$ where γ is the test threshold defined by a constraint on the false alarm rate $p(\lambda > \gamma|H_0) = p_{fa}$. Here p_{fa} is the specified false alarm rate level.

Under the null hypothesis λ is asymptotically χ^2-distributed with N complex degrees of freedom. The pdf of a chi-square distributed random variable with N complex degrees of freedom is given by

$$f(z) = \frac{1}{(N-1)!} z^{N-1} e^{-z},\ z > 0 \tag{27.15}$$

which is a gamma distribution with integer parameters N and 1. Consequently, the proposed test can be designed to satisfy a constraint on the tolerated false alarm rate while maximizing the detection probability, similarly to the Neyman–Pearson detector. This property provides a rigorous way of ensuring that a radar is not overwhelmed with tracking too many targets, or in a cognitive radio context rigorously controlling when idle spectrum may be accessed when the primary user is not active, for example.

Man-made signals typically exhibit cyclostationarity at multiple cyclic frequencies caused by symbol rate, coding and guard periods, and carrier frequency or their multiples. In order to take into account the rich information present in such signals, the non-parametric detector above may be easily extended to multiple cyclic frequencies as in Lunden et al. (2007).

The proposed spatial sign cyclic correlation detector has very low computational complexity. Moreover, the covariance matrix of the estimator has a very simple form under the null hypothesis, depending only on the number of samples M. In addition, employing the spatial sign cyclic correlation estimator allows for straightforward sequential computation of the test statistic as will be shown in the following section. This is obviously beneficial in detecting rapid changes in the observed signal statistics.

27.4.1 Sequential Test for Rapid Changes

Sequential test allows for making a decision without delay as soon as necessary evidence for making decisions at specified error levels is acquired. The presented sequential test is a truncated one. Truncation means that there is a maximum number of samples $M_{\max}$ that can be taken until a decision has to be made. The expression for

the test statistic in (27.14) and the spatial sign cyclic correlation estimator in (27.6) lend themselves to sequential calculation of the test statistic. Consequently, $\lambda(t) = \frac{1}{M_{\max}}||\hat{\boldsymbol{r}}_{S,t}(\alpha)||^2$ where the components of $\hat{\boldsymbol{r}}_{S,t}(\alpha)$ may be calculated sequentially

$$\hat{R}_{S,t}(\alpha,\tau) = \hat{R}_{S,t-1}(\alpha,\tau) + S(x(t))S(x^*(t+\tau))e^{-j2\pi\alpha t}, \tag{27.16}$$

for $t > 0$, and $\hat{R}_{S,0}(\alpha,\tau) = S(x(0))S(x^*(\tau))$.

The increments $S(x(t))S(x^*(t+\tau))e^{-j2\pi\alpha t}$ on the right-hand side of (27.16) are i.i.d. samples drawn from a uniform distribution on the unit circle under hypothesis H_0. Thus, $\hat{R}_{S,t}(\alpha,\tau)$ is a random walk in the complex plane (planar motion) with i.i.d. increments given by a uniform distribution on the unit circle.

Skipping the detailed derivation presented in Lunden et al. (2010), we can define a truncated sequential detection test as follows:

$$\begin{aligned} &\text{If } \lambda(t) \geq \gamma_s \text{ and } t \leq M_{\max}, && \text{Decide } H_1 \\ &\text{If } \lambda(t) < \gamma_s \text{ and } t = M_{\max}, && \text{Decide } H_0 \\ &\text{Otherwise,} && \text{Take a new sample, } t = t+1, \end{aligned} \tag{27.17}$$

where $\lambda(t) = \frac{1}{M_{\max}}||\hat{\boldsymbol{r}}_{S,t}(\alpha)||^2$, γ_s is the threshold for choosing among the alternatives, and $M_{\max}$ defines the truncation of the test, i.e., the maximum number of samples that can be acquired until a decision has to be made. Consequently, only a decision to accept H_1 can be made before the truncation point is reached. This strategy is suitable for minimizing the probability of missed detection. Under the null hypothesis $\lambda(M_{\max})$ is asymptotically chi-square distributed with N complex degrees of freedom where N is the number of lags. Consequently, the test threshold can be easily set based on the distributions of the test statistics and employed decision making policy (e.g., Neyman–Pearson). See Lunden et al. (2010) for details.

27.4.2 Decentralized Testing Using Multiple Detectors

In demanding applications such as radar, surveillance, and cognitive radio systems, multiple detectors may cooperate in order to improve the reliability of the test. Typical scenario where decentralized testing gives impressive gains is when signals experience shadowing and fading effects and interference. These effects can be mitigated by exploiting spatial diversity through collaboration among multiple detectors in different locations. In this section, we present a decentralized test for cyclostationarity.

A key assumption is that the test statistics of the secondary users are independent conditioned on hypotheses H_0 and H_1. The test statistics from individual detectors can be combined using the basic factorization principle for statistically independent

as follows:

$$\lambda_L = \sum_{i=1}^{L} \lambda^{(i)} \tag{27.18}$$

where L is the number of co-operating detectors and $\lambda^{(i)}$ denotes the spatial sign cyclic correlation test statistic of the ith individual detector. Since the test statistics $\lambda^{(i)}$ for each local detector is the local log-likelihoods under the null hypothesis, the sum of the single-user test statistics in (27.18) follows assuming conditional independence of the local test statistics.

The sum of independent gamma (or chi-square) distributed random variables is also gamma (or chi-square) distributed where the shape parameter of the distribution of the sum is the sum of the shape parameters of the distributions of the individual random variables. Hence, under the null hypothesis λ_L is chi-square distributed with LN complex degrees of freedom (i.e., gamma distributed with parameters LN and 1).

27.5 Simulation Examples

In this section the performance of the proposed single cycle spatial sign cyclic correlation based detector is studied in simulation. The detector is designed for one nonzero cyclic frequency. First, we illustrate that the spatial sign nonlinearity preserves the cyclic frequencies due to the symbol frequency for a cyclic prefix OFDM signal, see Lunden et al. (2010), Reed (1959), and Jacovitti and Neri (1994) for more detail. Figure 27.2 shows the normalized squared modulus of the cyclic correlation for a cyclic prefix OFDM signal for the conventional and spatial sign estimators. It can be seen that the cyclic frequencies are preserved by the spatial

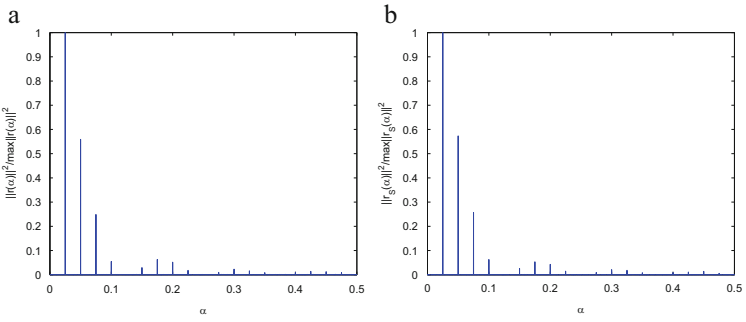

Fig. 27.2 Normalized squared modulus of (**a**) the cyclic correlation, and (**b**) the spatial sign cyclic correlation. The signal is a cyclic prefix OFDM signal with symbol frequency of 0.025. The spatial sign nonlinearity preserves the cyclic frequencies

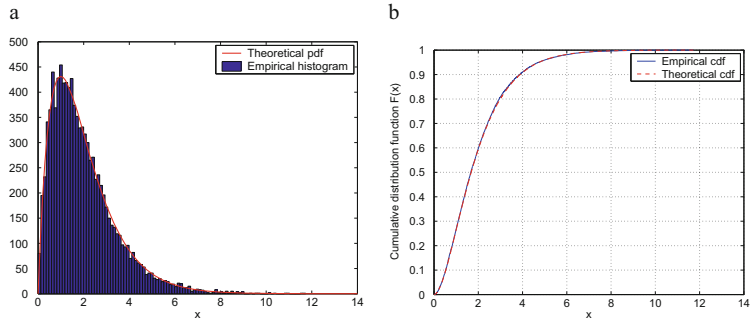

Fig. 27.3 Validity of the central limit theorem based approximation for the spatial sign cyclic correlation estimator. (**a**) Probability density function (pdf) and (**b**) cumulative distribution function (cdf) of the test statistic $\lambda = M||\hat{\boldsymbol{r}}_S(\alpha)||^2$ for a contaminated Gaussian noise distribution $0.95N_C(0, \sigma^2) + 0.05N_C(0, 25\sigma^2)$. The sample size $M = 100$. The measured empirical distribution is with high fidelity approximated by the distribution derived using the CLT

sign nonlinearity. Note that in practice the magnitudes of the cyclic correlation and spatial cyclic correlation are different. Magnitudes have been normalized here for easier comparison.

We will validate the result on the central limit theorem based approximation of the spatial sign cyclic correlation estimator by considering the distribution of the spatial sign cyclic correlation based test statistic. Figure 27.3 plots the probability density functions and the cumulative distribution functions of the test statistic $\lambda = M||\hat{\boldsymbol{r}}_S(\alpha)||^2$ for a complex, circularly symmetric i.i.d. contaminated Gaussian process $0.95N_C(0, \sigma^2) + 0.05N_C(0, 25\sigma^2)$. In this contamination model the noise comes with a probability of 0.95 from the nominal distribution $N_C(0, \sigma^2)$ and with a probability of 0.05 from the contaminating distribution $N_C(0, 25\sigma^2)$. The number of data samples is $M = 100$. The employed cyclic frequency was selected randomly from the interval $[0.05, 0.5]$ in the simulation. The histogram and empirical cumulative distribution functions have been obtained from 10,000 experiments. The measured empirical distribution is with very high fidelity approximated by the derived theoretical distribution, thus validating the central limit theorem based approximation employed in establishing the result.

The first test signal is an OFDM signal in DVB-T system ETSI (2004) with a fast-Fourier transform (FFT) size $N_{FFT} = 8192$ and a cyclic prefix of $N_{cp} = 1024$ samples. The total symbol length is given by $N_{FFT} + N_{cp}$. The number of subcarriers used is 6817, and 64-QAM (quadrature amplitude modulation) is used, and the length of the signal is three OFDM symbols (≈ 3 ms). The signal was sampled at the Nyquist rate. Thus, the oversampling factor wrt. the symbol rate is $N_{FFT} + N_{cp}$. Cyclic prefix OFDM signal induces cyclostationary with respect to the symbol frequency. Thus, the detection may be done at the symbol frequency by all detectors. In addition, all the detectors except SCD, which takes the weighted sum over all time delays, employ two time lags $\pm N_{FFT}$. The cyclic autocorrelation of the OFDM signal peaks for these time lags (Öner and Jondral 2007). In the

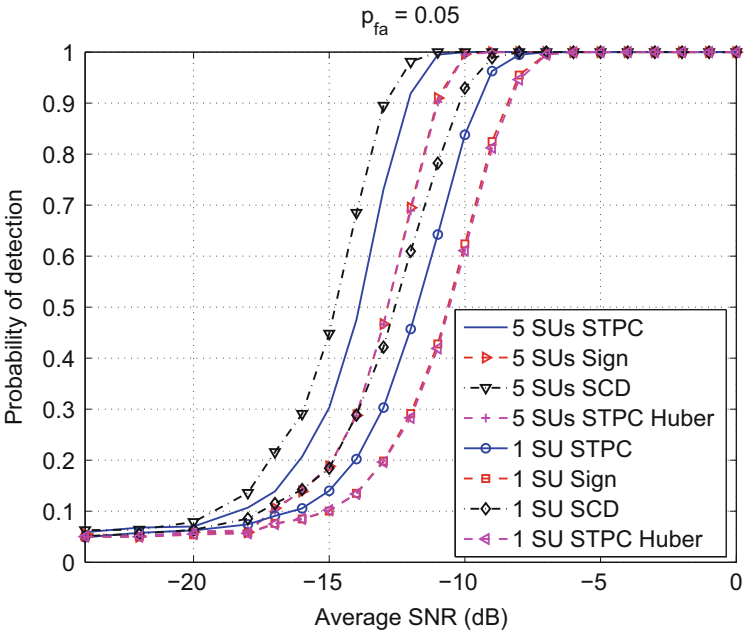

Fig. 27.4 Probability of detection vs. SNR (dB) in an AWGN channel for 1 and 5 detectors. The signal is an OFDM signal (DVB-T). Spatial sign cyclic correlation based detector suffers from a small performance degradation in comparison with the methods based on conventional cyclic correlation estimator in nominal Gaussian noise for the OFDM signal

comparisons we use the sub-optimum single-cycle detector (SCD) in Gaussian noise that correlates the ideal spectral correlation function (SCF) with the estimated SCF for one cyclic frequency of the signal (Gardner 1988); STPC (statistical test for the presence of cyclostationarity) method of Dandawaté and Giannakis (1994) and Lunden et al. (2007). We consider the cases where there is only one detector as well as the cooperative case where decision statistics from five co-operating detectors are combined.

Figure 27.4 plots the performance of the detectors for 1 and 5 co-operating detector in AWGN channel as a function of the Signal-to-Noise Ratio (SNR). The SNR is $\mathrm{SNR} = 10\log_{10}\frac{\sigma_x^2}{\sigma_n^2}$ where σ_x^2 and σ_n^2 denote the variances of the transmitted signal and the additive noise, respectively. False alarm rate constraint is set to 0.05. The same false alarm rate is employed in all the simulations. The simulation curves in the plots are averaged over 1000 independent experiments.

It can be seen that employing the spatial sign nonlinearity causes performance degradation in AWGN channel compared to the SCD and STPC detectors. The SCD detector has the best performance as expected since it requires more prior knowledge than any of the other detectors. However, this makes it also more sensitive to synchronization errors. The SCD methods, unlike the other detectors, require the knowledge of the exact carrier frequency, which is a nuisance parameter that needs

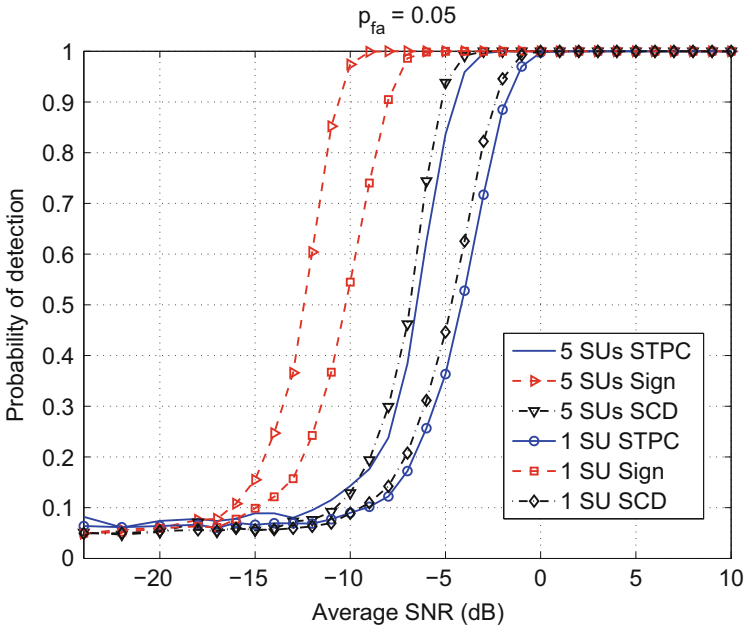

Fig. 27.5 Probability of detection vs. SNR (dB) in a contaminated additive Gaussian noise channel for 1 and 5 secondary users. Additive noise has a contaminated Gaussian distribution $0.95N_C(0, \sigma^2) + 0.05N_C(0, 100\sigma^2)$. SNR is defined with respect to σ^2. The signal is an OFDM signal (DVB-T). The spatial sign cyclic correlation based detector is more robust against heavy-tailed non-Gaussian noise than the other considered detectors

to be estimated in practice. In all the simulation examples, the carrier frequency was assumed to be known.

Figure 27.5 depicts the performance of the detectors in a contaminated Gaussian distribution. The contaminated Gaussian distribution is $0.95N_C(0, \sigma^2) + 0.05N_C(0, 100\sigma^2)$. The robustness of the spatial sign cyclic correlation based detector compared to the STPC detector and the SCD is clearly evident in the contaminated Gaussian noise case.

Figure 27.6 depicts the performance of the detectors in additive complex isotropic Cauchy distributed noise to demonstrate the highly robust performance of the proposed detectors. The noise location parameter is zero. Because the moments of Cauchy distribution are not defined, the performances are defined as a function of the generalized SNR (GSNR). The GSNR is defined as $\text{GSNR} = 10\log_{10}\frac{\sigma_x^2}{\gamma}$ where σ_x^2 is the variance of the transmitted signal and γ is the dispersion parameter of the Cauchy noise. The spatial sign cyclic correlation based detector is significantly more robust against the heavy-tailed Cauchy noise than the other detectors.

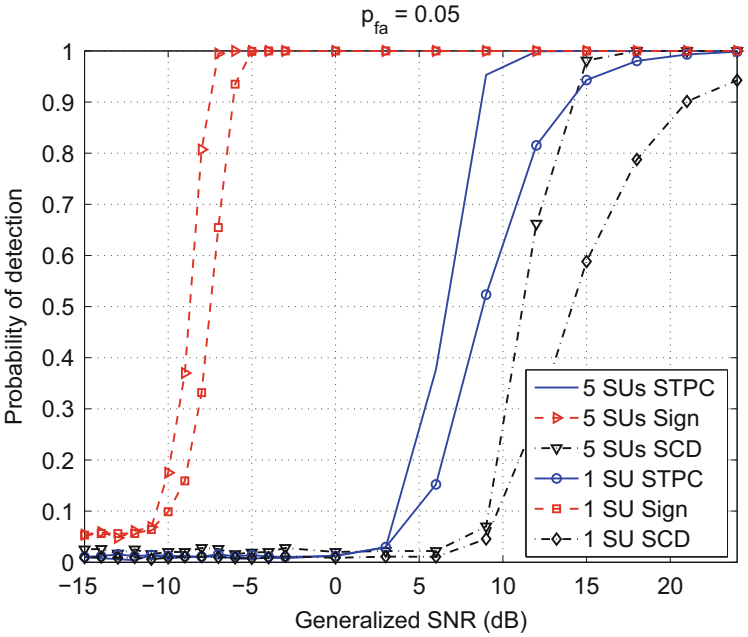

Fig. 27.6 Probability of detection vs. Generalized SNR (dB) in an additive complex isotropic Cauchy noise channel for 1 and 5 secondary users. The signal is an OFDM signal (DVB-T). The spatial sign cyclic correlation based detector is more robust against the heavy-tailed Cauchy noise than the other detectors

Figure 27.7 depicts the performance of the sequential test and the fixed sample size spatial sign detector as a function of the SNR. The channel is a frequency flat and Rayleigh fading. The transmitted signal is an IEEE 802.11 WLAN OFDM signal. The size of the FFT $N_{FFT} = 64$, the number of active subcarriers is 52, the length of the cyclic prefix $N_{cp} = 16$ samples, and QPSK modulation is employed. The maximum sensing time (truncation time) for the sequential detectors and also used by the fixed sample size detectors is 3 ms. The sequential detectors have comparable performance to the fixed sample size spatial sign cyclic correlation detectors. Note that for the sequential detectors the upper bound of the false alarm rate is 0.05. Figure 27.7 plots the average proportion of samples required to make a decision for the sequential detectors as a function of the SNR. The FC test refers to the minimum sensing time of the secondary users. Sequential detection schemes provide considerable reduction in the detection times.

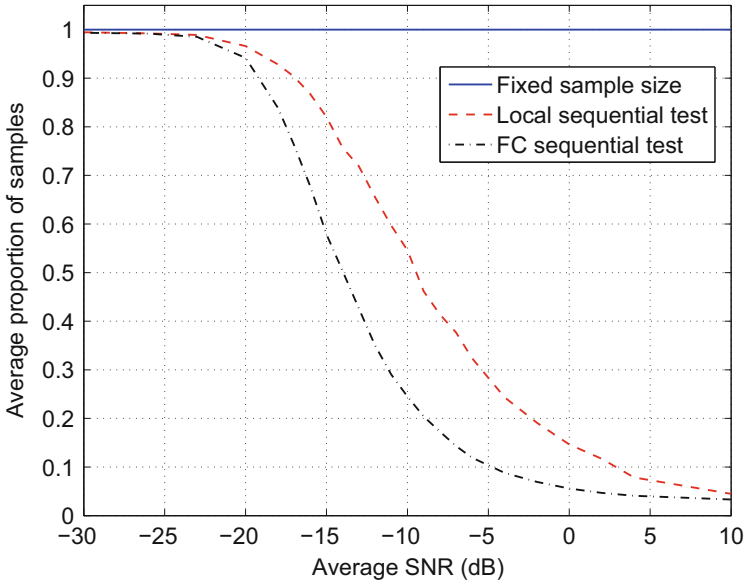

Fig. 27.7 Average proportion of samples used vs. Average SNR (dB) in a frequency flat Rayleigh fading channel for local and FC tests for a cooperation of five users. Additive noise has a Gaussian distribution. The signal is an OFDM signal (WLAN). The maximum sensing time is 3 ms. The sequential detection scheme reduces the required number of samples significantly compared to the fixed sample size spatial sign cyclic correlation detector with approximately equal performance

27.6 Conclusion

In this chapter a nonparametric spatial sign cyclic correlation based detector has been presented. It was shown that the spatial sign cyclic correlation function preserves the periodicity in autocorrelation function, hence facilitating the detection of cyclostationary signals. The distribution of the test statistics under the null hypothesis is derived. Tests for single and multiple co-operating detectors have been developed. Significant performance gains over the other considered methods for the non-Gaussian noise cases are obtained: 2–15 dB gain depending on the how heavy the tails of the noise distribution are for the OFDM and QPSK signals. Furthermore, a sequential spatial sign cyclic correlation based detection scheme was presented. It allows for detecting rapid changes and provides a shorter detection time on average than the fixed sample size detector with roughly equal performance.

References

Croux, C., Ollila, E., Oja, H.: Sign and rank covariance matrices: statistical properties and application to principal components analysis. In: Dodge, Y. (ed.) Statistical Data Analysis Based on the L1-norm and Related Methods, pp. 257–271. Birkhauser, Basel (2002)

Dandawaté, A.V., Giannakis, G.B.: Statistical tests for presence of cyclostationarity. IEEE Trans. Signal Process. **42**(9), 2355–2369 (1994)

European Telecommunications Standards Institute, Digital Video Broadcasting (DVB), Framing structure, channel coding and modulation for digital terrestrial television ETSI EN 300 744 v1.5.1 (2004–11), April 11, 2004. [Online]. Available: http://pda.etsi.org/pda/queryform.asp [Accessed Apr. 10, 2008]

Fraser, D.A.S.: Nonparametric Methods in Statistics. Wiley, New York (1957)

Gardner, W.A.: Signal interception: a unifying theoretical framework for feature detection. IEEE Trans. Commun. **36**(8), 897–906 (1988)

Jacovitti, G., Neri, A.: Estimation of the autocorrelation function of complex Gaussian stationary processes by amplitude clipped signals. IEEE Trans. Inf. Theory **40**(1), 239–245 (1994)

Kassam, S.A.: Signal Detection in Non-Gaussian Noise. Springer, New York (1988)

Kassam, S.A.: Nonparametric signal detection. In: Poor, H.V., Thomas, J.B. (Eds.) Advances in Statistical Signal Processing, vol. 2, pp. 66–91. JAI Press, Greenwich, CT (1993)

Lundén, J., Koivunen, V., Huttunen, A., Poor, H.V.: Spectrum sensing in cognitive radios based on multiple cyclic frequencies. In: Proceedings of the 2nd International Conference on Cognitive Radio Oriented Wireless Networks and Communications, pp. 37–43 (2007)

Lundén, J., Kassam, S.A., Koivunen, V.: Robust nonparametric cyclic correlation based spectrum sensing for cognitive radio. IEEE Trans. Signal Process. **58**(1), 38–52 (2010)

Middleton, D.: Non-Gaussian noise models in signal processing for telecommunications: new methods and results for Class A and Class B noise models. IEEE Trans. Inf. Theory **45**(4), 1129–1149 (1999)

Ollila, E., Tyler, D.E., Koivunen, V., Poor, H.V.: Complex elliptically symmetric distributions: survey, new results and applications. IEEE Trans. Signal Process. **60**(11), 5597–5625 (2012)

Öner, M., Jondral, F.: Air interface identification for software radio systems. Int. J. Electron. Commun. **61**(2), 104–117 (2007)

Reed, I.S.: On the use of Laguerre polynomials in treating the envelope and phase components of narrow-band Gaussian noise. IRE Trans. Inf. Theory **IT–5**, 102–105 (1959)

Sánchez, M.G., de Haro, L., Calvo, M., Mansilla, A., Montero, C., Oliver, D.: Impulsive noise measurements and characterization in a UHF digital TV channel. IEEE Trans. Electromagn. Compat. **41**(2), 124–136 (1999)

Visuri, S., Koivunen, V., Oja, H.: Sign and rank covariance matrices. J. Stat. Plann. Inference **91**(2), 557–575 (2000)