MARTIN T. WELLS
ASHIS SENGUPTA
Editors

Advances in Directional and Linear Statistics

A Festschrift for Sreenivasa Rao Jammalamadaka

Physica-Verlag
An Imprint of Springer Science+Business Media

Advances in Directional and Linear Statistics

Martin T. Wells • Ashis SenGupta
Editors

Advances in Directional and Linear Statistics

A Festschrift for Sreenivasa Rao Jammalamadaka

Physica-Verlag

Editors
Prof. Martin T. Wells
Cornell University
Statistical Science
Ithaca, NY 14853
USA
mtw1@cornell.edu

Prof. Ashis SenGupta
Indian Statistical Institute
Applied Statistics Unit
Barrackpore Trunk Road 203
700035 Kolkata
India
ashis@isical.ac.in

ISBN 978-3-7908-2627-2 e-ISBN 9978-3-7908-2628-9
DOI 10.1007/978-3-7908-2628-9
Springer Heidelberg Dordrecht London New York

Cover design: WMX Design GmbH, Heidelberg, Germany

Printed on acid-free paper

Physica-Verlag is a brand of Springer-Verlag Berlin Heidelberg
Springer-Verlag is part of Springer Science+Business Media (www.springer.com)

This volume is dedicated to Professor Sreenivasa Rao Jammalamadaka. Sreenivas has advanced statistical science in deep and original ways, and the statistics community has been enriched by his research contributions, intellectual leadership, and gracious mentorship of many junior colleagues.

Preface

Professor Sreenivasa Rao Jammalamadaka, formerly known as J.S. Rao and affectionately known to most of us as JS, was born on December 7, 1944, at Munipalle, India. Being under-aged for engineering studies was a blessing in disguise, and he was among the first batch of students selected for the Bachelor of Statistics (B. Stat.) degree at the Indian Statistical Institute (ISI), Kolkata. He received his Masters and Ph.D. degrees also from the ISI, and has the distinction of being the first B.Stat.-M.Stat.-Ph.D. of the ISI. He received his education from such legendary figures as Professors P.C. Mahalanobis, J.B.S. Haldane, C.R. Rao, and D. Basu among others, and worked with Professor C.R. Rao for his Ph.D. (1969) on path-breaking research in the then newly emerging area of directional data analysis.

JS moved to the USA in 1969 and was a faculty member at the Indiana University and then at the University of Wisconsin, Madison, before he moved to the University of California, Santa Barbara (UCSB) in 1976, where he has remained since then. At UCSB he played a pivotal and leading role in creating the new Department of Statistics and Applied Probability and was its first Chairman. He has been a prodigious mentor to the graduate students in that department, having provided guidance to as many of 35 Ph.D. students, at the last count. Throughout his career, JS has been extraordinarily generous to his colleagues in India, inviting them to the US and spending many of his sabbaticals helping Universities in India as well as in other countries.

JS has published extensively in leading international journals. His research work spans a wide spectrum which includes: goodness-of-fit tests, linear models, non-parametric and semi-parametric inference, Bayesian analysis, reliability, spacings statistics, and most notably, directional data analysis. He has written several books, both for undergraduate students as well as for advanced researchers. He has collaborated with a large number of researchers from around the world in general, and from India in particular. A Fellow of both the ASA and the IMS among other professional organizations, he has served nobly the cause of statistics at many national and international levels, including that of the President of the International Indian Statistical Association.

The present volume consists of papers written by students, colleagues and collaborators of JS from various countries, and covers a variety of research topics which JS enjoys and contributed immensely to. We are deeply thankful to these distinguished

authors for their contributions. We thank the anonymous referees for their reviews and suggestions. We also profoundly thank Dr. Kaushik Ghosh; his technical savvy, editorial contributions, and deep dedication to this project was extremely important in bringing this volume to fruition. Finally, it is a pleasure to acknowledge the understanding, patience and cooperation of Springer-Verlag in bringing out this volume. It is with great pleasure and pride that we present to our readers this volume, a Festschrift in celebration of the 65th birth anniversary of JS. We also take this opportunity to wish JS many, many more years of active research which will enrich the statistical sciences as well as foster further academic collaborations from far and wide.

Ithaca, NY *Martin T. Wells*
Kolkata, India *Ashis SenGupta*
June 2010

Contents

Contributors

Barry C. Arnold Department of Statistics, University of California, Riverside, CA 92521, USA, barry.arnold@ucr.edu

Debasis Bhattacharya Department of Statistics, University of California, Davis, CA 95616, USA, debasis_us@yahoo.com,dbhatta@wald.ucdavis.edu

Echarif Elharfaoui Département de Mathématiques et Informatique, Université Chouaîb Doukkali, Faculté des Sciences El Jadida, Rte Ben Maachou, B.P. 20, 24000, Maroc, and IMT (UMR CNRS 5219), Université Paul Sabatier, Toulouse Cedex 31062, France, elharfa@cict.fr

Frank J. Fabozzi Yale School of Management, New Haven, CT 06520-8200, USA, frank.fabozzi@yale.edu

Xiaoqi Gao Mathematics Department, University of Manchester, Alan Turing Building, Oxford Road, Manchester, M13 9PL, UK, Xiaoqi.Gao@postgrad.manchester.ac.uk

Riccardo Gatto Department of Mathematics and Statistics, University of Bern, 3012 Bern, Switzerland, gatto@stat.unibe.ch

Kaushik Ghosh Department of Mathematical Sciences, University of Nevada at Las Vegas, Box 454020, 4505 Maryland Parkway, Las Vegas, NV 89154-4020, USA, kaushik.ghosh@unlv.edu

Michel Harel IUFM du Limousin, Limoges Cedex 87036, France, and IMT (UMR CNRS 5219), Université Paul Sabatier, Toulouse Cedex 31062, France

Arvind Jammalamadaka Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA, ajamma@gmail.com

Yihua Jiang Capital One Financial Corp., 15000 Capital One Dr., Richmond, VA 23238, USA, Yihua.Jiang@capitalone.com

Peter Karcher Yello Strom GmbH, Taubenholzweg 1, 51105 Cologne, Germany, peter.karcher@gmx.de

S. Kasala Department of Mathematics and Statistics, University of North Carolina, Wilmington, NC 28403, USA, kasalas@uncw.edu

Young Shin Kim Department of Statistics, Econometrics and Mathematical Finance, School of Economics and Business Engineering, University of Karlsruhe and KIT, Geb. 20.12, 76131 Karlsruhe, Germany, kim@statistik.uni-karlsruhe.de

Subhash Kochar Department of Mathematics and Statistics, Portland State University, Portland, OR 97201, USA, kochar@pdx.edu

Tomasz J. Kozubowski Department of Mathematics and Statistics, University of Nevada, MS 084, Reno, NV 89557, USA, tkozubow@unr.edu

Vasudevan Mangalam Department of Mathematics, Universiti Brunei Darussalam, Gadong BE 1410, Brunei Darussalam, vmangalam@gmail.com

Saidbek S. Mirakhmedov Institute of Algorithm and Engineering, Fayzulla Hodjaev-45, Tashkent 700149, Uzbekistan, saeed_0810@yahoo.com

Sherzod M. Mirakhmedov Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi 23460, Pakistan, shmirakhmedov@yahoo.com

Christoph Möller Department of Statistics, Econometrics and Mathematical Finance, School of Economics and Business Engineering, University of Karlsruhe and KIT, Geb. 20.12, 76131 Karlsruhe, Germany, moeller@statistik.uni-karlsruhe.de

Anna K. Panorska Department of Mathematics and Statistics, University of Nevada, MS 084, Reno, NV 89557, USA, ania@unr.edu

Mathew D. Penrose Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK, m.d.penrose@bath.ac.uk

Madan L. Puri Department of Mathematics, Indiana University, Bloomington, IN 47405, USA
and
King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia, puri@indiana.edu

Fares Qeadan Department of Statistics and Probability, Michigan State University, A416 Wells Hall, East Lansing, MI 48823, USA, fares99us@yahoo.com

Svetlozar T. Rachev Econometrics and Mathematical Finance, School of Economics and Business Engineering, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
and
Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, USA, rachev@statistik.uni-karlsruhe.de

Bo Ranneby Centre of Biostochastics, Swedish University of Agricultural Sciences, 901 83, Umeå, Sweden, bo.ranneby@sekon.slu.se

B.L.S. Prakasa Rao Department of Mathematics and Statistics, University of Hyderabad, Hyderabad 500 046, India, blsprao@gmail.com

M. Bhaskara Rao Department of Environmental Health, Center for Genome Information, University of Cincinnati, Cincinnati, OH 45267, USA, raomb@ucmail.uc.edu

T. Subba Rao Mathematics Department, University of Manchester, Alan Turing Building, Oxford Road, Manchester, M13 9PL, UK
and
C.R. Rao Advanced Institute for Mathematics, Statistics and Computer Science, University of Hyderabad, Hyderabad, India, Tata.Subbarao@manchester.ac.uk

George G. Roussas Department of Statistics, University of California, Davis, CA 95616, USA, roussas@ucdavis.edu

Ashis SenGupta Applied Statistics Unit, Indian Statistical Institute, Kolkata 700035, India, ashis@isical.ac.in

György Terdik Department of Information Technology, Faculty of Informatics, University of Debrecen, Debrecen, Hungary, TERDIK@delfin.UNIDEB.HU

Yuedong Wang Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, USA, yuedong@pstat.ucsb.edu

Martin T. Wells Department of Statistical Science, Cornell University, 1190 Comstock Hall, Ithaca, NY 14853, USA, mtw1@cornell.edu

Maochao Xu Department of Mathematics and Statistics, Portland State University, Portland, OR 97201, USA, maochao@pdx.edu

Jun Yu Centre of Biostochastics, Swedish University of Agricultural Sciences, 901 83, Umeå, Sweden, jun.yu@sekon.slu.se

J.E. Yukich Department of Mathematics, Lehigh University, Bethlehem, PA 18015, USA, joseph.yukich@lehigh.edu

Dabao Zhang Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067, USA, zhangdb@stat.purdue.edu

H. Zhang Department of Mathematics and Statistics, Mississippi State University, Mississippi State, MS 39762, USA, hzhang@math.msstate.edu

Chapter 1
Models for Axial Data

Barry C. Arnold and Ashis SenGupta

Abstract A variety of models have been proposed to accommodate data involving directional vectors in two-dimensions, with no identified start or end point. By convention such directions are represented by points in the interval $(0, \pi)$. Data of this kind is called axial data. A survey of symmetric and asymmetric models for axial data is provided here. In addition, for certain models, parametric inference issues are addressed. In some cases, bivariate and multivariate extensions are readily envisioned.

1.1 Introduction

To model directions of vectors which do not have an identifiable start point or end point, an angular distribution with support $(0, \pi)$ is required. It is customary that such distributions are called axial distributions. Such axial data has often been analyzed by multiplying the observation values by two and then seeking a circular distribution (often a circular normal density) that provides an acceptable fit. This will be referred to as the "doubling" strategy, which is justifiable in many cases. Some times alternative approaches can yield better results. An alternative approach involves the recognition that in many cases axial variables can be viewed as circular variables modulo π. Arnold and SenGupta [1] proposed use of this approach, beginning with a circular normal density for the original directions. Wrapping strategies can be used to generate axial models as can approaches involving projections of points determined by a given bivariate distribution in $\mathcal{R}^2$. In the case of circular data, the most popular model is the von Mises or circular normal density. Its popularity can perhaps be traced to the fact that the model is a two parameter exponential family with minimal sufficient statistics that are simply sums of sines and cosines of the observed values. Only by a doubling approach can a similar convenient exponential

B.C. Arnold (✉)
Department of Statistics, University of California, Riverside, CA 92521, USA
e-mail: barry.arnold@ucr.edu

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_1,

family model be arrived at for axial data. This means that estimation and inference for almost all axial models will be "computer intensive" to some degree.

It is time to begin the tour of "the world of axial modeling".

1.2 Some Well-Known Circular Models

It will be useful to quickly review some of the more popular circular models (see e.g., [6]). The first in our list (and undoubtedly the most frequently encountered) is the circular normal distribution. A random variable Θ will be said to have circular normal distribution with parameters $\mu \in [0, 2\pi)$ and $\kappa \in (0, \infty)$ if its density is of the form

$$f_\Theta(\theta; \mu, \kappa) \propto \exp[\kappa(\theta - \mu)] I(0 \leq \theta < 2\pi). \tag{1.1}$$

The parameter μ corresponds to the mode of the density while κ is a precision parameter. If a random variable has density (1.1), we write $\Theta \sim CN(\mu, \kappa)$. An alternative parameterization is possible, as follows

$$f_\Theta(\theta; \alpha, \beta) \propto \exp[\alpha \cos\theta + \beta \sin\theta] I(0 \leq \theta < 2\pi). \tag{1.2}$$

In this version the parameters α and β can take on any real value. Representation (1.2), makes it evident that we are dealing with a two parameter exponential family of densities and that, if we have a sample of size n from this density, the complete minimal sufficient statistic will be

$$\left(\sum_{i=1}^n \cos\Theta_i, \sum_{i=1}^n \sin\Theta_i\right). \tag{1.3}$$

An alternative approach to the construction of models for circular data involves "wrapping" a real variable or equivalently evaluating that variable modulo 2π to yield a variable with support $[0, 2\pi)$. Thus we begin with a random variable X with support $(-\infty, \infty)$ and define the corresponding "wrapped" version of X by

$$\Theta = X \pmod{2\pi}. \tag{1.4}$$

This is clearly a rich source of models since the distribution of X can be quite arbitrary.

A popular choice for the distribution of the random variable to be "wrapped" is a symmetric stable distribution. In particular, the normal distribution or the Cauchy distribution is often selected. The wrapped normal distribution has a density expressible in the form of an infinite series, which complicates estimation and inference for the model. On the other hand, a wrapped Cauchy distribution has a density of the following relatively simple analytic form

$$f_\Theta(\theta; \alpha, \beta) \propto (1 + \alpha\cos\theta + \beta\sin\theta)^{-1} I(0 \leq \theta < 2\pi) \tag{1.5}$$

where α, β are real valued parameters satisfying $\alpha^2 + \beta^2 < 1$. Note that the choice $\alpha = \beta = 0$ yields the uniform distribution on $[0, 2\pi)$.

Another distribution which is sometimes considered for univariate directional data is the cardioid distribution with density of the form

$$f_\Theta(\theta; \alpha, \beta) \propto [1 + \alpha \cos\theta + \beta \sin\theta] I(0 \leq \theta < 2\pi) \tag{1.6}$$

in which α, β are real valued parameters satisfying $\alpha^2 + \beta^2 < 1$. A more general model, introduced by [7] will be called a generalized Cardioid density. It is of the form

$$f_\Theta(\theta; \alpha, \beta) \propto [1 + \alpha \cos\theta + \beta \sin\theta]^\gamma I(0 \leq \theta < 2\pi) \tag{1.7}$$

in which α, β, γ are real valued parameters satisfying $\alpha^2 + \beta^2 < 1$. The third parameter γ results in a flexible family of models which includes the wrapped Cauchy, the uniform and the cardioid distributions as special cases corresponding to $\gamma = -1, 0$ and 1 respectively.

An alternative flexible family of circular distributions are those of the "angular" form. For these , one begins with a arbitrary two-dimensional random vector (X, Y) with support in the plane. We then consider the angle in the interval $(0, 2\pi)$ determined by the line joining the origin to the point (X, Y), i.e., $\Theta = \tan^{-1}(Y/X)$. A popular choice for the distribution of (X, Y) in this construction is the classical bivariate normal distribution.

The last entry in our list of general methods for creating flexible circular models, is the technique of "restriction to the unit circle". In this approach one begins with an arbitrary joint density $f_0(x, y)$ whose support includes an open set that includes the unit circle. Then one considers the values taken on by this joint density at all points on the unit circle. Parameterizing the unit circle by $\theta \in [0, 2\pi)$, one obtains in this way a circular density of the form:

$$f(\theta) \propto f_0(\cos\theta, \sin\theta) I(0 \leq \theta < 2\pi). \tag{1.8}$$

In fact, in this construction f_0 can be replaced by a quite arbitrary function of two variables, not necessarily a joint density.

1.3 Introducing Asymmetry

Most of the circular models described in Sect. 1.2 are symmetric about their modes. But of course, many real world data configurations do not exhibit such symmetry. To accommodate this situation a variety of asymmetric versions of the densities have been proposed. We will return to discuss asymmetry in more detail in the context of axial distributions in Sect. 1.5. In the circular context, only a few brief comments will be presented.

First note that some of the constructions introduced in Sect. 1.2 clearly do not always produce symmetric models. Thus, restriction to the unit circle and angular constructions can easily yield asymmetric models. Likewise, wrapping an asymmetric density usually leads to asymmetry (examples in the literature include the wrapped exponential [5] and the wrapped skew-normal [9]).

Umbach and Jammalamadaka [10], motivated by [4] construction of skewed distributions on the real line, describe a general method of constructing skewed circular distributions. They begin with two symmetric circular densities f and g on $[-\pi, \pi)$. Denote the distribution function of g by G and choose w to be an odd function with period 2π and satisfying $|w(\theta)| \leq \pi$. They then verify that the function $2f(\theta)G(w(\theta))$ is an asymmetric circular density.

Efforts to develop asymmetric versions of the circular normal distribution often involve introduction of trigonometric functions of different frequencies. Thus, for example, we might consider a model of the form

$$f_\Theta(\theta; \alpha, \beta, \gamma, \delta) \propto \exp[\alpha \cos\theta + \beta \sin\theta + \gamma \cos(2\theta) + \delta \sin(3\theta)] I(0 \leq \theta < 2\pi). \tag{1.9}$$

Analogous adjustments can be made to the generalized cardioid density to obtain asymmetric versions.

1.4 Axial Models

As mentioned in the introduction, axial distributions are appropriate for modeling the directions of vectors without an identifiable beginning or end point. They typically will take on values between 0 and π. Two general strategies are considered for generating axial models from available circular models: wrapping and doubling. The doubling approach involves multiplying the axial variable by 2 and assuming that the doubled variable has one of the available circular distributions described in Sect. 1.2. Thus if our axial random variable is denoted by Θ and if $g(\theta)$ is a given circular density, then the density for Θ will be of the form

$$f_\Theta(\theta) = 2g(2\theta) I(0 \leq \theta < \pi). \tag{1.10}$$

Generally speaking such models will be suspect, because there seems to be little reason to expect that two times an axial variable will be sensibly modeled by a well known circular distribution. Nevertheless this technique does provide a very wide ranging flexible array of possible models which might be useful for modeling and/or prediction. There is an exceptional case in which doubling does seem to make sense, to be described later.

The second approach for generating axial models involves wrapping, but in this case wrapping around a half circle instead of a full circle as was done to generate circular models. Thus one begins with an arbitrary real valued random variable X and defines an axial variable Θ by setting

$$\Theta = X \pmod{\pi}. \tag{1.11}$$

Generally speaking it is desirable that axial density $f(\theta)$ should have the property that $\lim_{\theta \to \pi} f(\theta) = f(0)$, since the angles 0 and π are indistinguishable. In parallel fashion it is reasonable to expect that a circular density will satisfy $\lim_{\theta \to 2\pi} f(\theta) = f(0)$. Such restrictions will limit the class of acceptable densities for the random variable X that is to be wrapped.

Perhaps the most natural genesis for an axial variable involves envisioning a circular random variable that is imperfectly visible in the sense that any pair of values of the form θ and $\theta + \pi$ are indistinguishable. Thus if the original circular variable is denote by Φ the density of the observed axial variable will be of the form

$$f_{\Theta}(\theta) = [f_{\Phi}(\theta) + f_{\Phi}(\theta + \pi)]\, I(0 \leq \theta < \pi). \tag{1.12}$$

An alternative viewpoint, is that this density results from the fact that $\Theta = \Phi \pmod{\pi}$.

A particularly popular version of this construction begins with circular normal distribution for Φ and results in what is called an axial normal distribution for Θ of the form

$$f_{\Theta}(\theta) \propto \cosh(\alpha \sin\theta + \beta \cos\theta)\, I(0 \leq \theta < \pi). \tag{1.13}$$

See [1] for more detailed discussion of such axial normal variables. These axial normal variables, derived in a natural fashion from circular normal variables, are somewhat more difficult to deal with because the corresponding collection of densities does not comprise an exponential family.

In passing, it may be noted that beginning with a circular normal density and using the doubling approach does lead to an exponential family of axial densities of the form

$$f_{\Theta}(\theta; \alpha, \beta) \propto \exp[\alpha \cos 2\theta + \beta \sin 2\theta]\, I(0 \leq \theta < \pi). \tag{1.14}$$

The attraction of such models, with their conveniently available minimal sufficient statistic, is undeniable. Despite the lack of logical justification for use of the doubling approach in this context, it must be admitted that they comprise a flexible two parameter family of axial distributions which merits consideration for use in fitting data.

Instead of using the wrapping or the doubling approach, angular axial models may be considered. To this end we, as in the circular case, begin by considering a bivariate random vector (X, Y) with support in the plane. Then let Θ denote the angle in the interval $(0, \pi)$ determined by the extended line joining the origin to the point (X, Y). Thus $\Theta = \tan^{-1}(Y/X) + (\pi/2)$ where $\tan^{-1}$ is chosen to range over the interval $(-\pi/2, \pi/2)$ as is customary. Although the model is motivated by consideration of angles determined by a random point in the plane, it can be described in a slightly simpler fashion. The random vector (X, Y) only enters into the model in terms of its ratio Y/X and since the distribution of (X, Y) is arbitrary, so is the distribution of the real random variable $W = Y/X$. So the general angular

axial model can be represented in the form

$$\Theta = \tan^{-1} W + (\pi/2) \tag{1.15}$$

where W is an arbitrary real random variable.

A popular choice for the distribution of (X, Y) in this construction is the classical bivariate normal distribution with mean vector $\underline{\mu}$ and variance covariance matrix Σ. With this choice of distribution for (X, Y) the random variable $W = Y/X$ will have a Cauchy distribution and as $\underline{\mu}$ and Σ range over all possible values, the location and scale parameters of the Cauchy random variable W will range over all possible values, i.e., $-\infty < \mu < \infty$ and $0 < \sigma < \infty$.

With such a choice for the distribution of (X, Y), or equivalently of W, the resulting axial distribution is known as the angular central Gaussian distribution [8]. There are several possible parameterizations but the one providing the most simply interpretable parameters is of the form

$$f_\Theta(\theta) \propto \left(1 + \alpha \sin\left(2\left(\theta + \frac{\pi}{4} - \beta\right)\right)\right)^{-1} I(0 < \theta < \pi). \tag{1.16}$$

The parameter α controls the precision (or peakedness) of the density while β controls the location of the mode.

One obvious advantage of this model is that a simple transformation applied to the density yields a Cauchy density so that inference techniques appropriate for Cauchy distributed data may be directly used in analysis. But there is another, perhaps unexpected special property of the angular central Gaussian (ACG) distribution. These features of the model become more apparent when alternative parameterizations are considered.

A parameterization of the ACG density that highlights its linkage with a Cauchy(μ, σ) density is as follows:

$$f_\Theta(\theta) = \frac{2\sigma \;\; I(0 \le \theta < \pi)}{\pi(1 + \sigma^2 + \mu^2)[1 - \frac{2\mu}{1+\sigma^2+\mu^2} \sin(2\theta - \pi) - \frac{1-\mu^2-\sigma^2}{1+\sigma^2+\mu^2} \cos(2\theta - \pi)]} \tag{1.17}$$

where $-\infty < \mu < \infty$ and $0 < \sigma < \infty$.

Instead one might consider a version of the ACG density parameterized as follows

$$f_\Theta(\theta) = \frac{\sqrt{1 - \delta^2 - \gamma^2} \;\; I(0 \le \theta < \pi)}{\pi(1 + \gamma \sin(2\theta) + \delta \cos(2\theta))}. \tag{1.18}$$

where the parameters γ and δ are constrained by

$$0 \le \gamma^2 + \delta^2 < 1.$$

Of course the choice $\gamma = \delta = 0$ leads to a uniform distribution on $[0, \pi)$. If Θ has the density (1.18), it is clear that the doubled random variable Λ has a density of the form

$$f_{\Lambda)}(\lambda) \propto (1 + \gamma \sin\lambda + \delta \cos\lambda)^{-1}\ I(0 \leq \lambda < 2\pi), \tag{1.19}$$

which is recognizable as a wrapped Cauchy distribution (c.f., (1.5)). It would thus be possible to double data arising from an ACG distribution and analyze the doubled data using a wrapped Cauchy model.

Remark. The double linkage between the ACG distribution and the Cauchy distribution raises an interesting characterization problem. It is apparent from the above discussion that if W has a Cauchy$(0, 1)$ distribution, then for some a, b the random variables $U = a + bW \pmod{2\pi}$ and $V = 2(\tan^{-1} W + (\pi/2)) = 2\tan^{-1} W + \pi$ are identically distributed. Is this only true in the case in which W has a Cauchy distribution?

Parallel to the development of the generalized cardioid circular model, it is natural to extend the ACG distribution by introducing an additional parameter, say ρ, to consider generalized ACG densities of the form

$$f_{\Theta}(\theta) \propto [1 + \gamma \sin(2\theta) + \delta \cos(2\theta)]^{\rho}\ I(0 \leq \theta < \pi). \tag{1.20}$$

Yet another manner in which an axial density can be constructed is by "restriction to the unit half-circle". As in the analogous method of generating a circular density, one begins with an arbitrary joint density $f_0(x, y)$ whose support includes an open set that includes the unit half circle. Then one considers the values taken on by this joint density at all points on the unit half circle. Parameterizing the unit half circle by $\theta \in [0, \pi)$, one obtains in this way an axial density of the form:

$$f(\theta) \propto f_0(\cos\theta, \sin\theta) I(0 \leq \theta < \pi). \tag{1.21}$$

1.5 Asymmetric Axial Models

For axial distributions, as was the case for circular models, the majority of the popular densities are symmetric about their modal value. Asymmetric variations are however available. Wrapping an asymmetric real random variable X around the unit half circle will typically yield an asymmetric axial distribution. Such is also the case if we make use of the angular projection approach, i.e., consider $\Theta = \tan^{-1} W + (\pi/2)$ where W is an arbitrary real random variable. In this construction, the choice of a Cauchy distribution for W will lead to a symmetric density for Θ, but other choices for the distribution of W lead to asymmetric axial densities. Even a standard normal choice for W leads to asymmetry! An intriguing open problem is to characterize the class of distributions for W that will lead to symmetric axial distributions. It is not difficult to verify that this class includes t-densities with possibly fractional degrees of freedom, but presumably there are many more possibilities.

Finite mixtures of symmetric axial densities will typically exhibit asymmetry, as will finite logarithmic mixtures. More generally, if $f_1(\theta), f_2(\theta), \ldots, f_m(\theta)$

are m symmetric axial densities, then subject to non-negativity and integrability constraints one may consider the following usually asymmetric axial models

$$f(\theta) \propto \left[\sum_{i=1}^{m} a_i \ f_i(\theta)\right] / I(0 \le \theta < \pi), \tag{1.22}$$

and

$$f(\theta) \propto \left[\prod_{i=1}^{m} \{f_i(\theta)\}^{a_i}\right] / I(0 \le \theta < \pi). \tag{1.23}$$

It is possible to develop asymmetric versions of the axial normal distribution by the introduction of trigonometric functions of different frequencies. Thus, for example, one might consider a density of the form

$$f_\Theta(\theta; \alpha, \beta, \gamma, \delta) \propto \cosh[\alpha \sin\theta + \beta \cos\theta + \gamma \sin(2\theta) + \delta \cos(3\theta)] I(0 \le \theta < \pi) \tag{1.24}$$

Analogous adjustments can be made to the ACG and extended ACG model to obtain asymmetric versions such as

$$f_\Theta(\theta) \propto [1 + \gamma \sin(2\theta) + \delta \cos(2\theta)) + \lambda \sin(4\theta) + \tau \cos(6\theta)]^\rho \ I(0 \le \theta < \pi). \tag{1.25}$$

1.6 Bivariate Axial Distributions

Bivariate (and multivariate) versions of the axial distribution (1.15) are readily constructed by choosing (W_1, W_2) to have a bivariate (or $\underline{W}$ to have a multivariate) Cauchy distribution with Cauchy$(0, 1)$ marginals and then applying a transformation of the form (1.15) to each coordinate, i.e., define

$$\Theta_i = \tan^{-1}(\mu_1 + \sigma_1 W_1) + (\pi/2)$$

for each i.

For some applications it may be desirable to have a bivariate (or multivariate) axial distribution with conditional densities belonging to specific axial families.

For example, if we wish to deal with bivariate distributions with conditionals in the generalized ACG family (1.20), we may consider densities of the form

$$f_{\Theta_1,\Theta_2}(\theta_1, \theta_2) \propto \left[(1, \sin 2\theta_1, \cos 2\theta_1) A (1, \sin 2\theta_2, \cos 2\theta_2)'\right]^\rho \\ I(0 \le \theta_1 < \pi, 0 \le \theta_2, \pi). \tag{1.26}$$

It is evident that (1.1) does indeed have all conditional densities of Θ_1 given Θ_2 and of Θ_2 given Θ_1 in the family (1.20). The choice $\rho = -1$, yields densities with ACG conditionals.

Similarly, it is possible to describe a bivariate density with all conditionals in the axial normal family (1.13). It will be of the form;

$$f(\theta_1, \theta_2) \propto \cosh[(\sin\theta_1, \cos\theta_1) B (\sin\theta_2, \cos\theta_2))'] \; I(0 \leq \theta_1 < \pi, 0 \leq \theta_2 < \pi), \tag{1.27}$$

where B is a 2×2 matrix of parameters.

In the case of axial model (1.14), the fact that this is two parameter exponential family allows us (using [3]) to identify the class of all bivariate axial densities with conditional densities of the form (1.14). Such densities will be of the form

$$\begin{aligned} f(\theta_1, \theta_2) = \exp[(1, \sin 2\theta_1, \cos 2\theta_1) A (1, \sin 2\theta_2, \cos 2\theta_2)'] \\ I(0 \leq \theta_1 < \pi, 0 \leq \theta_2, \pi), \end{aligned} \tag{1.28}$$

where A is a 3×3 real matrix in which a_{11} is a normalizing constant that is a function of the other a_{ij}'s so that the density will integrate to 1. Multivariate versions of this construction can be analogously defined.

Asymmetric multivariate axial models can be constructed using basically the same devices that were used to construct asymmetric univariate axial models. More details on these constructions may be found in [2].

References

1. Arnold B, SenGupta A (2006) Probability distributions and statistical inference for axial data. Environ Ecol Stat 12:271–285
2. Arnold B, SenGupta A (2009) Flexible bivariate circular models. In: Advances in multivariate statistical methods. World Scientific, Singapore
3. Arnold B, Strauss D (1991) Bivariate distributions with conditionals in prescribed exponential families. J R Stat Soc Series B 53:365–375
4. Azzalini A (2005) The skew-normal distribution and related multivariate families. Scand J Stat 32:159–200
5. Jammalamadaka SR, Kozubowski T (2004) New families of wrapped distributions for modeling skew circular data. Commun Stat Theory Methods 33(9):2059–2074
6. Jammalamadaka SR, SenGupta A (2001) Topics in circular statistics. World Scientific, Singapore
7. Jones MC, Pewsey A (2005) A family of symmetric distributions on the circle. J Am Stat Assoc 100:1422–1428
8. Mardia KV, Jupp PE (2000) Directional statistics. Wiley, New York
9. Pewsey A (2006) Modeling asymmetrically distributed circular data using the wrapped skew-normal distribution. Environ Ecol Stat 13:257–269
10. Umbach D, Jammalamadaka SR (2009) Building asymmetry into circular distributions. Stat Probab Lett 79:659–663

Chapter 2
Asymptotic Behavior of the Universally Consistent Conditional U-Statistics for Nonstationary and Absolutely Regular Processes

Echarif Elharfaoui, Michel Harel, and Madan L. Puri

Abstract A general class of conditional U-statistics was introduced by W. Stute as a generalization of the Nadaraya–Watson estimates of a regression function. It was shown that such statistics are universally consistent. Also, universal consistencies of the window and k_n-nearest neighbor estimators (as two special cases of the conditional U-statistics) were proved. Later, (Harel and Puri, Ann Inst Stat Math 56(4):819–832, 2004) extended his results from the i.i.d. case to the absolute regular case. In this paper, we extend these results from the stationary case to the nonstationary case.

2.1 Introduction

Let $\{Z_i = (X_i, Y_i);\ i \in \mathbb{N}^*\}$ be a sequence of random vectors with continuous distribution functions $H_i(z)$, $i \in \mathbb{N}^*$, $z \in \mathbb{R}^d \times \mathbb{R}^s$, defined on some probability space $(\Omega, \mathcal{A}, P)$.

Assume that H_i admits a strictly positive density and H_i has the two marginals F_i and G_i.

Let h be a function of k-variates (the U kernel) such that for some $r > 2$, $h \in \mathcal{L}_r^*$, which means that $E\{\sup_\beta |h(Y_\beta)|^r\} < +\infty$ (where sup extends over all permutations $\beta = (\beta_1, \ldots, \beta_k)$ of length k, that is, over all pairwise distinct $\beta_1, \ldots, \beta_k$ taken from $\mathbb{N}^*$) which implies that for all integers $i_1, i_2, \ldots, i_k$ $(i_1 < i_2 < \ldots < i_k)$ $h(Y_{i_1}, \ldots, Y_{i_k}) \in \mathcal{L}_r$, the space of all random variables Z for which $|Z|^r$ is integrable. In order to measure the impact of a few X's, say $(X_1, \ldots, X_k)$, on a

E. Elharfaoui (✉)
Département de Mathématiques et Informatique, Université Chouaîb Doukkali, Faculté des Sciences El Jadida, Rte Ben Maachou, B.P. 20, 24000, Maroc, and IMT (UMR CNRS 5219), Université Paul Sabatier, Toulouse Cedex 31062, France
e-mail: elharfa@cict.fr

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_2,

function $h(Y_1, \ldots, Y_k)$ of the pertaining Y's, set

$$m(\mathbf{x}) \equiv m(x_1, \ldots, x_k) := E[h(Y_1, \ldots, Y_k) | X_1 = x_1, \ldots, X_k = x_k] \tag{2.1}$$

where m is defined on $\mathbb{R}^{dk}$.

For estimation of $m(\mathbf{x})$, [7] proposed a statistic of the form

$$u_n(\mathbf{x}) = u_n(x_1, \ldots, x_k) = \frac{\sum_\beta h(Y_{\beta_1}, \ldots, Y_{\beta_k}) \prod_{j=1}^k K[(x_j - X_{\beta_j})/h_n]}{\sum_\beta \prod_{j=1}^k K[(x_j - X_{\beta_j})/h_n]} \tag{2.2}$$

where u_n is defined on $\mathbb{R}^{dk}$, K is the so-called smoothing kernel satisfying $\int K(u)\, du = 1$ and $\{h_n,\ n \geq 1\}$ is a sequence of bandwidth tending to zero at appropriate rates. Here summation extends over all permutations $\beta = (\beta_1, \ldots, \beta_k)$ of length k, that is, over all pairwise distinct $\beta_1, \ldots, \beta_k$ taken from $1, \ldots, n$. Stute [7] proved the asymptotic normality and weak and strong consistency of $u_n(\mathbf{x})$ when the random variables $\{(X_i, Y_i),\ i \geq 1\}$ are independent and identically distributed. Harel and Puri [3] extended the results of [7] from independent case to the case when the underlying random variables are absolutely regular. Stute [9] also derived the $\mathcal{L}_r$ convergence of the conditional U-statistics under the i.i.d. set up.

If a number of the X_i's in the random sample are exactly equal to x which can happen if X is a discrete random variable, $P^Y(\cdot|X = x)$ can be estimated by the empirical distribution of the Y_i's corresponding to X_i's equal to x. If few or none of the X_i's are exactly equal to x, it is necessary to use Y_i's corresponding to X_i's near x. This leads to estimators $\widehat{P}_n^Y(\cdot|X = x)$ of the form

$$\widehat{P}_n^Y(\cdot|X = x) = \sum_{i=1}^n W_{ni}(x)\, \mathbb{1}_{[Y_i \in \cdot]}$$

where $W_{ni}(x) = W_{ni}(x, X_1, \ldots, X_n)$ $(1 \leq i \leq n)$ weights those values of i for which X_i is close to x more heavily than these values of i for which X_i is far from x and $\mathbb{1}_A$ denotes the indicator function of A.

Let g be a Borel function on $\mathbb{R}^s$ such that $g(Y) \in \mathcal{L}_r$. Corresponding to W_n is the estimator $l_n(x)$ of $l(x) = E(g(Y)|X = x)$ defined by

$$l_n(x) = \sum_{i=1}^n W_{ni}(x) g(Y_i).$$

More generally if we now consider the estimates of $m(x)$ defined in (2.4), this leads to weighting those values of β for which $\mathbf{X}_\beta = (X_{\beta_1}, \ldots, X_{\beta_k})$ is close to $\mathbf{x}$ more heavily than the values of β for which $\mathbf{X}_\beta$ is far from $\mathbf{x}$.

This is why, as in [8], we study a fairly general class of conditional U-statistics of the form

$$m_n(\mathbf{x}) = \sum_{\beta} W_{\beta,n}(\mathbf{x}) h(\mathbf{Y}_\beta) \tag{2.3}$$

designed to estimate $m(\mathbf{x})$, where $W_{\beta,n}(\mathbf{x})$ is defined from a function $W_n(\mathbf{x}, \mathbf{y})$ by $W_{\beta,n}(\mathbf{x}) = W_n(\mathbf{x}, \mathbf{X}_\beta)$, $\mathbf{Y}_\beta = (Y_{\beta_1}, \ldots, Y_{\beta_k})$, and the summation in (2.3) takes place over all permutations $\beta = (\beta_1, \ldots, \beta_k)$ of length k such that $1 \leq \beta_i \leq n$, $i = 1, \ldots, k$.

Remark 2.1. The estimator defined in (2.2) is a special case of the estimator defined in (2.3), see (2.23).

In order to make $m_n(\mathbf{x})$ a local average, $W_{\beta,n}(\mathbf{x})$ has to give larger weights to those $h(\mathbf{Y}_\beta)$ is close to $\mathbf{x}$. For this general class of conditional U-statistics (defined in (2.3)) and for i.i.d. random variables, [8] derived the universal consistency. Harel and Puri [4] extended his results from the i.i.d. case to the absolute regular case. In this paper, we extend it to the nonstationary case and absolutely regular r.v.'s which allow broader applications that include, among others, hidden Markov models (HMM) described in detail in [4].

We shall call $W_{\beta,n}$ *universally consistent* if and only if

$$m_n(\mathbf{X}) \to m(\mathbf{X}) \text{ in } \mathcal{L}_r$$

under no conditions on h (up to integrability) or the distribution of $\{(X_i, Y_i),\ i \geq 1\}$. Here $\mathbf{X} = (X_1^0, \ldots, X_k^0)$ is a vector of X's with the same distribution as $(X_1, \ldots, X_k)$ and independent of $\{(X_i, Y_i),\ i \geq 1\}$.

2.2 Preliminaries

Let $(Z_i)_{i \geq 1}$ be a stochastic process indexed by the positive integers, taking value in a finite dimensional Euclidean space $\mathbb{H}$. Identifying $\mathbb{H}$ with a product of a finite number copies or the real line, we write H_i for the distribution function of Z_i. We will assume that the process has some form of asymptotic stationarity, implying that the sequence H_i converges in a sense to be made precise to a limiting distribution function H.

For $i \leq j$, let $\mathcal{A}_i^j$ denote the σ-algebra of events generated by $Z_i, \ldots, Z_j$. We shall say that the nonstationary stochastic process is absolutely regular if

$$\sup_{n \in \mathbb{N}^*} \max_{1 \leq j \leq n-k} E \left\{ \sup_{A \in \mathcal{A}_{j+k}^{\infty}} \left| P(A \mid \mathcal{A}_1^j) - P(A) \right| \right\} = \beta(k)^\star \downarrow 0 \text{ as } n \to \infty$$

where $\mathbb{N}^* = \{1, 2, \ldots\}$.

All along the paper, we assume that $(\star)$ holds with a geometrical rate;

$$\sum_{m\geq 1} m\beta^{\frac{\delta}{1+\delta}}(m) < \infty \quad \text{for some } \delta > 0. \tag{2.4}$$

We consider a parameter ξ in $\mathbb{H}$ whose components can be naturally estimated by U-statistics. To be more formal and precise, we assume that ξ is defined as follows. Let k be an integer, to be the degree of the U-statistics. Let ϕ be a function from $\mathbb{H}^k$ into $\mathbb{H}$, invariant by permutation of its arguments. We are interested in parameters of the form

$$\xi = \int_{\mathbb{H}^k} \phi \, \mathrm{d}H^{\otimes k} = \int_{\mathbb{H}^k} \phi(z_1, \ldots, z_k) \prod_{l=1}^{k} dH(z_l). \tag{2.5}$$

and the function ϕ is called the kernel of the parameter ξ.

Example 2.1. Take $\mathbb{H}$ to be $\mathbb{R}$. The mean vector corresponds to taking $k = 1$ and ϕ is the identity.

Example 2.2. Take $\mathbb{H}$ to be $\mathbb{R}^2$. Consider ξ to be the two-dimensional vector whose components are the marginal variances. We take $k = 2$ and ϕ is going to be a function defined on $(\mathbb{R}^2)^2$. It has two arguments, each being in $\mathbb{R}^2$, and it is defined by

$$\phi\left((u, v), (u', v')\right) = \left(\frac{u^2 + u'^2}{2} - uu', \frac{v^2 + v'^2}{2} - vv'\right).$$

Such a parameter can be estimated naturally by a U-statistics, essentially replacing $H^{\otimes k}$ in (5) by an empirical counterpart. By using the invariance of ϕ, the estimator of ξ is then of the form

$$\widehat{\xi}_n = \binom{n}{k}^{-1} \sum_{\beta} \phi(Z_{\beta_1}, \ldots, Z_{\beta_k}). \tag{2.6}$$

To specify our assumption on the process, it is convenient to introduce copies of $\mathbb{H}$. Hence we write $\mathbb{H}_i$, $i \geq 1$, an infinite sequence of copies of $\mathbb{H}$. The basic idea is to think of the process at time i as taking value in $\mathbb{H}_i$ and we think of each $\mathbb{H}_i$ as the ith component of $\mathbb{H}^\infty$. We then agree on the following definition.

Definition 2.1. A canonical p-subspace of $\mathbb{H}^\infty$ is any subspace of the form $\mathbb{H}_{i_1} \oplus \ldots \oplus \mathbb{H}_{i_p}$ with $1 \leq i_1 < \cdots < i_p$. We write $\mathbb{S}_p$ for a generic canonical p-subspace.

Remark 2.2. For $(i_1, \ldots, i_p) \neq (j_1, \ldots, j_p)$, if we note $\mathbb{S}_p = \mathbb{H}_{i_1} \oplus \ldots \oplus \mathbb{H}_{i_p}$ and $\mathbb{S}'_p = \mathbb{H}_{j_1} \oplus \ldots \oplus \mathbb{H}_{j_p}$, we have $\mathbb{S}_p \neq \mathbb{S}'_p$, with $\mathbb{S}_p \subset \mathbb{H}^\infty$ and $\mathbb{S}'_p \subset \mathbb{H}^\infty$.

The origin of this terminology is that when $\mathbb{H}$ is the real line, then a canonical p-subspace is a subspace spanned by exactly p distinct vectors of the canonical

basis of $\mathbb{H}^\infty$. We write $\sum_{\mathbb{S}_p \subset \mathbb{H}^n}$ for a sum over all canonical p-subspaces included in $\mathbb{H}^n$.

To such a canonical subspace $\mathbb{S}_p = \mathbb{H}_{i_1} \oplus \ldots \oplus \mathbb{H}_{i_p}$ we can associate the distribution function $H_{\mathbb{S}_p}$ of $(Z_{i_1}, \ldots, Z_{i_p})$ as well as the distribution function with the same marginals

$$H^{\otimes \mathbb{S}_p} = \otimes_{1 \le j \le p} H_{i_j} = \otimes_{\mathbb{H}_i \subset \mathbb{S}_p} H_i. \tag{2.7}$$

Clearly the marginal of $H^{\otimes \mathbb{S}_p}$ are independent, while that of $H_{\mathbb{S}_p}$ are not.

Consider two nested canonical subspace $\mathbb{S}_p$ and $\mathbb{S}_{k-p}$ where $\mathbb{S}_{k-p} \subset \mathbb{H}^n \ominus \mathbb{S}_p$. For a function ϕ symmetric in its argument and defined on $\mathbb{S}_p \oplus \mathbb{S}_{k-p}$, we can define its projection onto the functions defined on $\mathbb{S}_p$ by

$$z \in \mathbb{S}_p \to \phi(z, \mathbb{S}_{k-p}) = \int_{\mathbb{S}_{k-p}} \phi(z, y)\, \mathrm{d}H^{\otimes \mathbb{S}_{k-p}}(y). \tag{2.8}$$

Identifying $\mathbb{S}_p \oplus \mathbb{S}_{k-p}$ with $\mathbb{H}^k$ and $\mathbb{H}^p$ with $\mathbb{S}_p$, that allows to project functions defined on $\mathbb{H}^k$ onto functions on $\mathbb{H}^p$. However, with this identification, the projection depends on the particular choice of $\mathbb{S}_{k-p}$ in $\mathbb{H}^n$. To remove the dependence in $\mathbb{S}_{k-p}$, we sum over all choices of $\mathbb{S}_{k-p}$ in $\mathbb{H}^n \ominus \mathbb{S}_p$ by

$$\phi_{\mathbb{S}_p}(z) = \binom{n-p}{k-p}^{-1} \sum_{\mathbb{S}_{k-p} \subset \mathbb{H}^n \ominus \mathbb{S}_p} \phi(z, \mathbb{S}_{k-p}). \tag{2.9}$$

Let k be an integer and for each $n \ge k$, consider a kernel $\phi_n \equiv \phi$ of degree k depending on n.

A U-statistics of degree k is defined by

$$U_n = \binom{n}{k}^{-1} \sum_{\beta} \phi_n(Z_{\beta_1}, \ldots, Z_{\beta_k}), \tag{2.10}$$

we can then define an analogue of Hoeffding decomposition when the random variables come from a nonstationary process. For this purpose, consider, firstly, an expectation of U_n if the process had no dependence, namely,

$$U_{n,0} = \binom{n}{k}^{-1} \sum_{\mathbb{S}_k \subset \mathbb{H}^n} \int_{\mathbb{S}_k} \phi\, \mathrm{d}H^{\otimes \mathbb{S}_k}. \tag{2.11}$$

Then for any $p = 1, \ldots, k$, we define

$$U_{n,p} = \binom{n}{p}^{-1} \sum_{\mathbb{S}_p \subset \mathbb{H}^n} \int_{\mathbb{S}_p} \phi_{\mathbb{S}_p}\, \mathrm{d} \otimes_{\mathbb{H}_i \subset \mathbb{S}_p} (\delta_{Z_i} - H_i) \tag{2.12}$$

where $\delta_{\{.\}}$ is the Dirac function.

Finally, for $p > k$, we set

$$U_{n,p} = 0. \tag{2.13}$$

The analogue of Hoeffding decomposition is the equality

$$U_n = \sum_{0 \le p \le k} \binom{k}{p} U_{n,p}. \tag{2.14}$$

Note that this decomposition makes an explicit use of convention (2.13), and this is why this convention was introduced.

We now need to specify exactly what we mean by asymptotic stationary of a process. For this, recall the following notion of distance between probability measures.

Definition 2.2. The distance in total variation between two probability measures P and Q defined on the same σ-algebra $\mathcal{A}$ is

$$|P - Q|_{\mathcal{A}} = \sup_{A \in \mathcal{A}} |P(A) - Q(A)|.$$

If $\mathbb{S}_p$ is a canonical subspace of $\mathbb{H}^\infty$, we write $\sigma_{\mathbb{S}_p}$ the σ-algebra generated by the Z_i's with $\mathbb{H}_i \subset \mathbb{S}_p$. We write P the probability measure pertaining to the process $(Z_i)_{i \ge 1}$, which is a probability measure on $\mathbb{H}^\infty$.

Definition 2.3. The process $(Z_i)_{i \ge 1}$ with probability measure P on $\mathbb{H}^\infty$ is geometrically asymptotically stationary if there exists a strictly stationary process with distribution Q on $\mathbb{H}^\infty$, and a positive τ less than 1, such that for $i \ge 1$,

$$|P - Q|_{\sigma_{\mathbb{H}_i}} \le \tau^i. \tag{2.15}$$

We suppose that there exists a strictly stationary process $(Z_i^*)_{i \ge 1}$ with probability measure Q on $\mathbb{H}^\infty$, which is absolutely regular with the same rate as the process $(Z_i)_{i \ge 1}$. H is the distribution function of Z_i^*, H admits a strictly positive density and H has the two marginals F and G.

We define the function ϕ^* on $\mathbb{H}_1$ by

$$z \in \mathbb{H}_1 \mapsto \phi^*(z, \mathbb{H}^k \ominus \mathbb{H}_1) = \int_{\mathbb{H}^k \ominus \mathbb{H}_1} \phi(z, y)\, \mathrm{d}H^{\otimes(k-1)}. \tag{2.16}$$

Next, we denote

$$U_{n,1}^* = n^{-1} \sum_{i=1}^{n} \int_{\mathbb{H}_1} \phi^* d(\delta_{Z_i^*} - H).$$

2.3 Assumptions and Main Results

In this section, we identify $\mathbb{H} = \mathbb{H}' \times \mathbb{H}''$ with $\mathbb{R}^d \times \mathbb{R}^s$. For a generic canonical p-subspace S_p of $\mathbb{H}^\infty$, we write $S_{1,k}$ and $S_{2,k}$ its projections respectively in $\mathbb{H}'^\infty$ and $\mathbb{H}''^\infty$.

We consider the nonstationary sequence of random vectors $\{Z_i = (X_i, Y_i);\ i \in \mathbb{N}^*\}$ with values in $\mathbb{R}^d \times \mathbb{R}^s$ and continuous distribution functions H_i and H_i has the two marginals F_i and G_i.

We assume that the sequence $\{Z_i\}_{i\geq 1}$ is absolute regular with rates (2.4) and (2.15) is satisfied with its associated stationary sequence $\{\widetilde{Z}_i = (\widetilde{X}_i, \widetilde{Y}_i);\ i \in \mathbb{N}^*\}$ of stationary random vectors.

For the ease of convenience, we shall write W_β for $W_{\beta,n}$.

Consider the following set of assumptions:

(i) There exists functions $V_n(\mathbf{x}, \mathbf{y})$ on $\mathbb{R}^{2dk}$ such that for each $l \in \mathcal{L}_r^*$, $\mathbf{z}^{(n)} = (z_1, \ldots, z_n) \in \mathbb{R}^{dn}$ and $\mathbf{y}^{(n)} = (y_1, \ldots, y_n) \in \mathbb{R}^{sn}$

$$\sum_\beta W_n(\mathbf{x}, \mathbf{z}_\beta) l(\mathbf{y}_\beta) = \frac{\sum_\beta V_n(\mathbf{x}, \mathbf{z}_\beta) l(\mathbf{y}_\beta)}{\sum_\beta V_n(\mathbf{x}, \mathbf{z}_\beta)}$$

where $\mathbf{z}_\beta = (z_{\beta_1}, \ldots, z_{\beta_k})$ and $\mathbf{y}_\beta = (y_{\beta_1}, \ldots, y_{\beta_k})$.

(ii) There exists a function $V(\mathbf{x})$ on $\mathbb{R}^{dk}$ such that for each scalar function q on $\mathbb{R}^{dk}$ verifying

$$\int |V(\mathbf{x}) q(\mathbf{x})|^r dF^{\otimes k}(\mathbf{x}) < \infty$$

we have

$$\lim_{n\to\infty} \binom{n}{k}^{-1} \sum_{S_{1,k} \subset \mathbb{I}'^n} \int_{S_{1,k}} q(\mathbf{z}) V_n(\mathbf{x}, \mathbf{z}) dF^{\otimes S_{1,k}}(\mathbf{z}) = q(\mathbf{x}) \tilde{f}(\mathbf{x}) \int V(\mathbf{z}) d\mathbf{z}$$

where $\tilde{f}(\mathbf{x}) = \prod_{j=1}^k f(x_j)$ and f is the density function of F.

(iii) Define the kernel of degree k by

$$\phi_n(\mathbf{z}, \mathbf{y}) = h(\mathbf{y}) V_n(\mathbf{x}, \mathbf{z}) \Big/ \int V_n(\mathbf{x}, \mathbf{u}) dF^{\otimes k}(\mathbf{u}).$$

Suppose that

$$\sup_{\mathbb{S}_k \subset \mathbb{H}^\infty} \int_{\mathbb{S}_k} |\phi_n|^{2+2\delta}\, dP_{\sigma_{\mathbb{S}_k}} < \infty \tag{2.17}$$

$$\sup_{\mathbb{S}_k \subset \mathbb{H}^\infty} \int_{\mathbb{S}_k} |\phi_n|^{2+2\delta}\, dQ_{\sigma_{\mathbb{S}_k}} < \infty \tag{2.18}$$

where $\delta > 0$.

Remark 2.3. Our conditions (i) and (ii) are completely different from conditions (ii) to (v) in [8]. Our conditions are more general and more easy to verify. More, the condition (i) in [8] is not necessary.

The following theorems generalize Theorems 2.1, 2.2, 2.3 and 2.4 in [4] from the stationary dependent case to the nonstationary dependent case.

Theorem 2.1. *Assume that $h \in \mathcal{L}_r^*$. Then under (i)–(iii), (2.4) and (2.15),*

$$m_n(\mathbf{X}) \to m(\mathbf{X}) \text{ in } \mathcal{L}_r, \text{ as } n \to \infty,$$

where $r = 2 + 2\delta$, that is

$$E\left[|m_n(\mathbf{x}) - m(\mathbf{x})|^r \mu(d\mathbf{x})\right] \to 0 \text{ as } n \to \infty \tag{2.19}$$

where μ denotes the distribution of $(X_1, X_2, \ldots, X_k)$.

Corollary 2.1. *Under the conditions of Theorem 2.1 and (2.34) in Sect. 2.4, $m_n(\mathbf{x}) \to m(\mathbf{x})$ with probability one for μ-almost all $\mathbf{x}$.*

Remark 2.4. In [4], we supposed that h is bounded, this condition is not necessary now.

Theorems 2.2 and 2.3 deal with two special cases: window weights and *NN*-weights. Consistency of window estimates for the regression function has been obtained by [2] and [5]. *NN*-weights for the regression function have been studied in [6], Theorem 2.

In what follows, $|\cdot|$ denotes the maximum norm on $\mathbb{R}^d$. We also write

$$\|\mathbf{X}_\beta - \mathbf{x}\| := \max_{1 \le i \le k} |X_{\beta_i} - x_i|.$$

To define window weights, put (see [8])

$$W_\beta(\mathbf{x}) = \begin{cases} \mathbb{1}_{[\|\mathbf{X}_\beta - \mathbf{x}\| \le h_n]} / \sum_\beta \mathbb{1}_{[\|\mathbf{X}_\beta - \mathbf{x}\| \le h_n]} & \text{if well defined} \\ 0 & \text{otherwise.} \end{cases} \tag{2.20}$$

Here $h_n > 0$ is a given window size to be chosen by the statistician. Then we have the following results:

Theorem 2.2. *Assume $h_n \to 0$ and $nh_n^d \to \infty$ as $n \to \infty$. Then, under the conditions (2.4) and (2.15), we have*

$$m_n(\mathbf{X}) \to m(\mathbf{X}) \text{ in } \mathcal{L}_r,$$

where $W_\beta(\mathbf{x})$ in (2.3) is given by (2.20).

For the *NN*-weights, recall that X_j is among the $k_n NN$ of $x \in \mathbb{R}^d$ if $d_j(x) := \|X_j - x\|$ is among the k_n-smallest ordered values $d_{1:n}(x) \le \ldots \le d_{n:n}(x)$ of the d's. Ties may be broken by randomization.

For a given $1 \le k_n \le n$, set

$$W_\beta(\mathbf{x}) = \begin{cases} k_n^{-d} & \text{if } X_{\beta_i} \text{ is among the } k_n - NN \text{ of } x_i \text{ for } 1 \le i \le k \\ 0 & \text{otherwise.} \end{cases} \tag{2.21}$$

Theorem 2.3. *Assume that $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. Then, under the conditions (2.4) and (2.15), we have*

$$m_n(\mathbf{X}) \to m(\mathbf{X}) \ \text{ in } \mathcal{L}_r,$$

where $W_\beta(\mathbf{x})$ in (2.3) is given by (2.21).

We now consider as estimator of $m(x)$, the statistics of the form

$$m_n(\mathbf{x}) = u_n(\mathbf{x}) \tag{2.22}$$

where $u_n(\mathbf{x})$ is defined in (2.2). Then, in view of (2.3) we have

$$W_{\beta,n}(\mathbf{x}) = \frac{\prod_{j=1}^{k} K[(x_j - X_{\beta_j})/h_n]}{\sum_\beta \prod_{j=1}^{k} K[(x_j - X_{\beta_j})/h_n]} \tag{2.23}$$

where $K(\mathbf{x})$ is a so-called smoothing kernel satisfying $\int K(\mathbf{u})d\mathbf{u} = 1$ and $\lim_{\mathbf{u}\to\infty} |\mathbf{u}|K(\mathbf{u}) = 0$ and $\{h_n,\ n \geq 1\}$ is a sequence of bandwidths tending to zero. This special case was studied by [7] for i.i.d. random variables, and from Theorem 2.1, we can generalize his result for nonstationary dependent random variables. The following theorem establishes that the universal consistency still holds for conditional U-statistics involving kernel K and a sequence of bandwidth h_n.

Theorem 2.4. *Assume that $h_n \to 0$ and $nh_n^d \to \infty$ as $n \to \infty$. Then, under the conditions (2.4) and (2.15), we have*

$$m_n(\mathbf{X}) \to m(\mathbf{X}) \ \text{ in } \mathcal{L}_r,$$

where $m(\boldsymbol{x})$ is given (2.1).

2.4 Proof of Theorems and Corollary 2.1

First, we show that m_n is the ratio of two U-statistics. Let $\mathbf{x} = (x_1, \ldots, x_k)$ be fixed throughout. Let

$$U_n(h, \mathbf{x}) = U_n(\mathbf{x}) = U_n = \binom{n}{k}^{-1} \sum_\beta h(\mathbf{Y}_\beta) V_n(\mathbf{x}, \mathbf{X}_\beta) \Big/ \int V_n(\mathbf{x}, \mathbf{u}) dF^{\otimes k}(\mathbf{u}).$$

Hence $m_n(\mathbf{x}) = U_n(h, \mathbf{x}) \big/ U_n(1, \mathbf{x})$ and $U_n(h, \mathbf{x})$, for each $n \geq k$, is a nonstationary U-statistic as defined in (2.10) with a hind depending on n.

Consider the sequence of functionals

$$\theta_n(h, \mathbf{x}) \equiv \theta_n = \binom{n}{k}^{-1} \sum_{S_k \subset \mathbb{H}^n} \int_{S_k} \phi_n dH^{\otimes S_k}$$

where ϕ_n is defined in (iii).

Note that $\theta_n = E(U_n)$.

The decomposition defined in (2.14) can be written as

$$U_n = \theta_n + \sum_{p=1}^{k} \binom{k}{p} U_{n,p}$$

where $U_{n,p}$ is defined as in (2.12).

To prove Theorem 2.1, the following lemmas are needed.

Lemma 2.1. *Under the conditions of Theorem 2.1*

$$E(U_{n,p})^2 = \mathcal{O}(n^{-2}).$$

Proof. We shall consider the case $p = 2$. The proofs in the cases $c = 3, \ldots, k$ are analogous and so they are omitted.

We first note that

$$U_{n,2} = \binom{n}{2}^{-1} \sum_{S_2 \subset \mathbb{H}^n} \int_{S_2} \phi_{S_2} d_{\otimes_{\mathbb{H}_i \subset S_2}} (\delta_{Z_i} - H_i)$$

so we have

$$E(U_{n,2})^2 = \binom{n}{2}^{-1} \sum_{1 \le i_1 < i_2 \le n} \sum_{1 \le l_1 < l_2 \le n} J((i_1, i_2), (l_1, l_2))$$

where

$$J((i_1, i_2), (l_1, l_2)) = E\left(\int_{S_2} \phi_{S_2} d_{\otimes_{1 \le j \le 2} \mathbb{H}_{i_j}} (\delta_{Z_{i_j}} - H_{i_j}) \right.$$
$$\left. \times \int_{S_2'} \phi_{S_2'} d_{\otimes_{1 \le m \le 2} \mathbb{H}_{l_m}} (\delta_{Z_{l_m}} - H_{l_m}) \right)$$

$$S_2 = \mathbb{H}_{i_1} \oplus \mathbb{H}_{i_2} \text{ and } S_2' = \mathbb{H}_{l_1} \oplus \mathbb{H}_{l_2}.$$

So from condition (2.4) and condition (iii), we have from Lemma 2.1 in [10] the inequalities:

(i) If $1 \le i_1 < i_2 \le l_1 < l_2$, then

$$J((i_1, i_2), (l_1, l_2)) \le M\{\beta(l_2 - l_1)\}^{\frac{\delta}{1+\delta}};\ l_2 - l_1 \ge i_2 - i_1 \tag{2.24}$$

and

$$J((i_1, i_2), (l_1, l_2)) \le M\{\beta(i_2 - i_1)\}^{\frac{\delta}{1+\delta}};\ i_2 - i_1 \ge l_2 - l_1 \tag{2.25}$$

where M is a finite positive constant.

Thus, using (2.24) and (2.25), we obtain

$$\left| \sum_{1 \le i_1 < i_2 \le l_1 < l_2 \le n} J((i_1, i_2), (l_1, l_2)) \right| = \mathcal{O}(n^2). \tag{2.26}$$

Similarly
(ii) If $1 \le i_1 < l_1 < i_2 < l_2 \le n$, then

$$\left| \sum_{1 \le i_1 < l_1 < i_2 < l_2 \le n} J((i_1, i_2), (l_1, l_2)) \right| = \mathcal{O}(n^2). \tag{2.27}$$

(iii) If $1 \le i_1 < l_1 \le l_2 < i_2 \le n$, then

$$\left| \sum_{1 \le i_1 < l_1 \le l_2 < i_2 \le n} J((i_1, i_2), (l_1, l_2)) \right| = \mathcal{O}(n^2). \tag{2.28}$$

From (2.26), (2.27) and (2.28), we obtain

$$E(U_{n,2})^2 = \mathcal{O}(n^{-2}).$$

Thus the result for the case $p = 2$ is proved. □

Lemma 2.2. *Under the condition of Theorem 2.1, for μ-almost all $\boldsymbol{x}$*

$$\frac{\theta_n(h, \boldsymbol{x})}{\theta_n(1, \boldsymbol{x})} \to m(\boldsymbol{x}),\ n \to \infty.$$

Proof. By definition, we have

$$\theta_n(h, \mathbf{x}) \equiv \theta = \binom{n}{k}^{-1} \sum_{S_k \subset \mathbb{H}^n} \int_{S_k} \phi_n d\,\mathbb{H}^{\otimes S_k}.$$

Put

$$\theta_n'(h,\mathbf{x}) = \binom{n}{k}^{-1} \sum_{S_k \subset \mathbb{H}'^n} \int_{S_k} h(\mathbf{y}) V_n(\mathbf{x},\mathbf{z}) dF^{\otimes S_k}(\mathbf{y},\mathbf{z}).$$

From condition (2.15), we deduce

$$|\theta_n(h,\mathbf{x}) - \theta_n'(h,\mathbf{x})| \to 0 \ \text{ as } n \to \infty. \tag{2.29}$$

From condition (ii) in Sect. 2.3, we have

$$\lim_{n\to\infty} \binom{n}{k}^{-1} \sum_{S_{1,k} \subset \mathbb{H}'^n} \int_{S_{1,k}} m(\mathbf{z}) V_n(\mathbf{x},\mathbf{z}) dF^{\otimes S_{1,k}}(\mathbf{z}) = m(\mathbf{x}) \tilde{f}(\mathbf{x}) \int V(\mathbf{z}) d\mathbf{z} \tag{2.30}$$

and so

$$\lim_{n\to\infty} \binom{n}{k}^{-1} \sum_{S_{1,k} \subset \mathbb{H}'^n} \int_{S_{1,k}} V_n(\mathbf{x},\mathbf{z}) dF^{\otimes S_{1,k}}(\mathbf{z}) = \tilde{f}(\mathbf{x}) \int V(\mathbf{z}) d\mathbf{z}. \tag{2.31}$$

By definition, we have

$$\begin{aligned}
\theta_n(h,\mathbf{x}) &= \binom{n}{k}^{-1} \sum_{S_k \subset \mathbb{H}'^n} \int_{S_k} h(\mathbf{y}) V_n(\mathbf{x},\mathbf{z}) dF^{\otimes S_k}(\mathbf{y},\mathbf{z}) \\
&= \binom{n}{k}^{-1} \sum_{S_k \subset \mathbb{H}'^n} \int_{S_{1,k}} \left(\int_{S_{2,k}} E(h(\mathbf{y})|\mathbf{X}=\mathbf{z}) V_n(\mathbf{x},\mathbf{z}) dG^{\otimes S_{2,k}}(\mathbf{y}) \right) \\
&\quad \times dF^{\otimes S_{1,k}}(\mathbf{z}) \\
&= \binom{n}{k}^{-1} \sum_{S_{1,k} \subset \mathbb{H}'^n} \int_{S_{1,k}} m(\mathbf{z}) V_n(\mathbf{x},\mathbf{z}) dF^{\otimes S_{1,k}}(\mathbf{z}).
\end{aligned} \tag{2.32}$$

From (2.29)–(2.32), we deduce easily that

$$\frac{\theta_n(h,\mathbf{x})}{\theta_n(1,\mathbf{x})} \to m(\mathbf{x}), \ n \to \infty.$$

To prove Theorem 2.1, from Lemmas 2.1 and 2.2, we now have to show that for μ-almost all $\mathbf{x}$,

$$U_{n,1}(h,\mathbf{x}) \to 0 \text{ in probability.}$$

Since

$$U_{n,1}(h,\mathbf{x}) = n^{-1}\sum_{S_1\subset\mathbb{H}^n}\int_{S_1}\phi_{S_1}d\otimes_{\mathbb{H}_i\subset S_1}(\delta_{Z_i}-H_i)$$
$$= n^{-1}\sum_{i=1}^{n}\int_{\mathbb{H}_i}\phi_{\mathbb{H}_i}d(\delta_{Z_i}-H_i)$$

we have

$$E(U_{n,1})^2 = n^{-2}E\left(\sum_{i=1}^{n}\int_{\mathbb{H}_i}\phi_{\mathbb{H}_i}d(\delta_{Z_i}-H_i)\right)^2$$
$$= n^{-2}\sum_{i=1}^{n}E\left(\int_{\mathbb{H}_i}\phi_{\mathbb{H}_i}d(\delta_{Z_i}-H_i)\right)^2$$
$$+ 2n^{-2}\sum_{1\le i<j\le n}E\left\{\left(\int_{\mathbb{H}_i}\phi_{\mathbb{H}_i}d(\delta_{Z_i}-H_i)\right)\right.$$
$$\left.\times\left(\int_{\mathbb{H}_j}\phi_{\mathbb{H}_j}d(\delta_{Z_j}-H_j)\right)\right\}.$$

From Lemma 2.1 of [10] and condition (iii), we have

$$E(U_{n,1})^2 \le 2n^{-2}nM(2,h) + 4n^{-2}M^{\frac{1}{1+\delta}}(r,h)\sum_{p=1}^{n}(p+1)\beta^{\frac{\delta}{1+\delta}}(p)$$
$$= \mathcal{O}(n^{-1})$$

where $M(t,h) = E\{\sup_\beta|\phi_n(\mathbf{X}_\beta,\mathbf{Y}_\beta)|^t\}$, which implies

$$E(U_{n,1})^2 = \mathcal{O}(n^{-1}). \tag{2.33}$$

From Lemmas 2.1 and 2.2 and from (2.33), we have

$$U_n(h,\mathbf{x}) \to m(\mathbf{x})\tilde{f}(\mathbf{x})\int V(\mathbf{z})d\mathbf{z}$$

and

$$U_n(1,\mathbf{x}) \to \tilde{f}(\mathbf{x})\int V(\mathbf{z})d\mathbf{z} \text{ in probability}$$

as $n\to\infty$ for μ-almost all $\mathbf{x}$.
It remains to prove the uniform integrability.

It is an easy convergence of the Jensen's inequality

$$\begin{aligned}
&\sup_{n\in\mathbb{N}^*} E\left\{\left[\sum_{\beta} V_n(\mathbf{X},\mathbf{X}_\beta)|h(\mathbf{Y}_\beta)|/\sum_{\beta} V_n(\mathbf{X},\mathbf{X}_\beta)\right]^r\right\}\\
&\quad\le \sup_{n\in\mathbb{N}^*} E\left\{\sum_{\beta} V_n(\mathbf{X},\mathbf{X}_\beta)|h(\mathbf{Y}_\beta)|^r/\sum_{\beta} V_n(\mathbf{X},\mathbf{X}_\beta)\right\}\\
&\quad\le E\left\{\sup_{\beta}|h(\mathbf{Y}_\beta)|^r\right\} < +\infty.
\end{aligned}$$

and Theorem 2.1 is proved.

The proof of Corollary 2.1 is a consequence of Lemma 2.1 and Lemma 2.3 below.

For a d-dimensional vector V, consider the norm $\|V\| = \max_{1\le j\le d}|V^{(j)}|$. This norm is equivalent to the Euclidian norm and easy to work with here. We will use this norm in Lemma 2.3 below see also [1].

Lemma 2.3. *Let $(V_n)_{n\ge1}$ be a sequence of d-dimensional centered absolutely regular and non necessarily stationary random vectors with rate satisfying*

$$\sum_{i\ge1}(i)^{\frac{r-\delta}{2}}[\beta(i)]^{\frac{\delta}{r}} < \infty \tag{2.34}$$

$$\sup_{i\ge1} E(\|V_i\|^r) < \infty. \tag{2.35}$$

Then

$$n^{-1}\sum_{i=1}^{n} V_i \to 0 \ \text{ with probability 1, as } n\to\infty.$$

Proof. For $\epsilon > 0$,

$$\begin{aligned}
P\left(\frac{1}{n}\sum_{i=1}^{n}\|V_i\| \ge \epsilon\right) &= P\left(\max_{1\le j\le d}\left|\frac{1}{n}\sum_{i=1}^{n} V_i^{(j)}\right| \ge \epsilon\right)\\
&\le \sum_{1\le j\le d} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} V_i^{(j)}\right| \ge \epsilon\right).
\end{aligned} \tag{2.36}$$

For all $1 \le j \le d$, one has from Markov's inequality that

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} V_i^{(j)}\right| \ge \epsilon\right) \le \frac{1}{\epsilon^r n^r} E\left(\left|\sum_{i=1}^{n} V_i^{(j)}\right|^r\right). \tag{2.37}$$

By Lemma 5.2 of [3], one has that

$$E\left(\left|\sum_{i=1}^{n} V_i^{(j)}\right|^r\right) \leq Cn^{r/2}. \tag{2.38}$$

From the above two inequalities, one deduces that

$$E\left(\left|\sum_{i=1}^{n} V_i^{(j)}\right|^r\right) \leq \frac{C}{\epsilon^r} n^{r/2}. \tag{2.39}$$

Since $r/2 > 1$, the last inequality implies that for all $1 \leq j \leq d$,

$$\sum_{n\geq 1} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} V_i^{(j)}\right| \geq \epsilon\right) < \infty$$

which, in turn, implies that

$$\sum_{n\geq 1} P\left(\left\|\frac{1}{n}\sum_{i=1}^{n} V_i\right\| \geq \epsilon\right) < \infty.$$

Lemma 4.3 then follows by Borel–Cantelli theorem.

The proofs Theorems 2.2 to 2.4 are also consequences of Theorem 2.1 by using technics similar as in the proofs of Theorem 2.2 to Theorem 2.4 in [4]: that is to verify that conditions (i)–(iii) are satisfied.

References

1. Bennett G (1962) Probability inequalities for the sum on independent random variables. J Am Stat Assoc 19:33–45
2. Devroye LP, Wagner TJ (1980) Distribution-free consistency results in nonparametric discrimination and regression function estimate. Ann Stat 8:231–239.
3. Harel M, Puri LM (1996) Conditional U-statistics for dependent random varaibles. J Multivar Anal 57:84–100
4. Harel M, Puri LM (2004) Universally consistent conditional U-statistics for absolutely regular processes and its applications for hidden markov models. Ann Inst Stat Math 56(4):819–832
5. Speigelman G, Sacks J (1980) Consistent window estimation inonparametric regression. Ann Stat 8:240–246
6. Stone G (1977) Consistent nonparametric regression, Ann Stat 5:595–620
7. Stute W (1991) Conditional U-statistics. Ann Probab 19:812–825
8. Stute W (1994a) Universally consistent conditional U-statistics. Ann Stat 22:460–473
9. Stute W (1994b) L^p convergence of conditional U-statistics. J Multivar Anal 51:71–82
10. Yoshihara KL (1976) Limiting behavior of U-statistics for stationary absolutely regular processes. Z Wahrsch Verw Gebiete 35:237–252

Chapter 3
Regression Models with STARMA Errors: An Application to the Study of Temperature Variations in the Antarctic Peninsula

Xiaoqi Gao and T. Subba Rao

Abstract Motivated by spatio-temporal problems that occur in many areas such as environment, geography etc., we propose multivariate regression models with space-time ARMA errors. The methods of model identification and estimation are studied. The asymptotic properties of the estimators have been derived and simulations are provided. The methodology is applied to the analysis of monthly mean surface temperatures at five locations in the Antarctic Peninsula. This area of Antarctic is of great concern to climatologists, as it is believed that there is higher rate of warming compared to the rest. During the period from January 1978 to December 1998, the temperatures at all the five locations in the Antarctic Peninsula have increased. Substantial warming were detected at Faraday/Vernadsky and Rothera stations with the warming rate of 1.07 °C/decade and 1.08 °C/decade respectively and both trends are significant at level 1%.

3.1 Introduction

The Intergovernmental Panel on Climate Change (IPCC) confirmed that the global average surface temperature has increased by 0.6 ± 0.2 °C since the late 19th century (see [4]). Though the earth is unequivocally warming, the geographical patchiness of the recent climate change is evident since it has been reported that the Antarctic Peninsula (the northern most part of the mainland of Antarctica) is one of the rapid warming areas in recent 50 years (see [4]). Since the temperature records at industrial areas may be masked by both the increased concentrations of short-lived sulphate aerosols and the heating of urban meteorological stations, it has been suggested by [13] that the temperature trends of Antarctic stations are a particularly important indicator of climate change. Furthermore, the warming of the

X. Gao (✉)
Mathematics Department, University of Manchester, Alan Turing Building, Oxford Road, Manchester, M13 9PL, UK
e-mail: Xiaoqi.Gao@postgrad.manchester.ac.uk

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_3,

coastal areas in the Antarctic will lead to a collapse of ice shelves which results in an increase of sea level. Therefore, the analysis of temperatures at the Antarctic Peninsula is critical for understanding the dynamic behavior of climate change.

There are a number of recent studies concerned with the temperature trends of the Antarctic Peninsula (see, e.g., [12, 13]). These studies agree that the Antarctic Peninsula is warming and the statistically significant warming happened at Faraday/Vernadsky and Orcadas stations with the rates of 0.5 °C/decade and 0.2 °C/decade respectively. These trends were obtained by fitting a simple least square regression model with independent errors. Since the temperature series is obviously temporally autocorrelated, the significance of the trend using the adjusted t-test [9, 11] is not appropriate.

It is widely believed that besides temporal autocorrelation in the temperatures at each location, there is a strong correlation between spatial locations (see [5]). In time series modeling, these facts need to be taken into account. In view of this, the temperatures at different locations in the Antarctic Peninsula is obviously a spatio-temporal problem. Based on the belief that we will have a better insight into the dynamics by including spatial information in modeling rather than analyzing each series separately, we propose a multivariate regression model with correlated errors having spatio-temporal structure which takes into account both spatial dependencies and temporal autocorrelations. Several approaches have been proposed for the analysis of spatio-temporal processes but we prefer space-time autoregressive moving average (STARMA) models (see [7]) since these models are parsimonious and have been widely used for the data analysis in many areas of research, such as meteorology [9], criminology [7] and air pollution [1] etc.

The data we used in this paper is described in Sect. 3.2. Section 3.3 introduces the multivariate regression models with STARMA errors. The model estimation, selection of orders, and the diagnostic checking are also included in this section. The asymptotic properties and simulations of the estimators are considered in Sects. 3.4 and 3.5. The application of this model to the analysis of the temperatures in the Antarctic Peninsula is considered in Sect. 3.6. The results are summarized in Sect. 3.7.

3.2 Data

The Scientific Committee on Antarctic Research (SCAR), an inter-disciplinary committee of the International Council for Science (ICSU), established a Reference Antarctic Data for Environmental Research project which produced a database called READER. The database is posted on the following website: http://www.antarctica.ac.uk/met/READER/

The database consists of monthly mean surface temperatures of several Antarctic meteorological stations for recent decades. Temperatures are recorded in Celsius scale. The monthly mean data are computed using the mean of the 6-hourly synoptic observations. The longitude and latitude of each station are also provided by

READER. The details of the database and the quality control of the observations is discussed by [12].

There are some missing values in the database. Each missing value is estimated by taking the average of previous two observations and next two observations in the same month of this location. For example, the estimate of the missing value in June 1995 can be obtained by taking the average of the observations in June 1993, June 1994, June 1996 and June 1997. To obtain reliable data, we only consider continuous periods without many missing values.

The database includes records from research stations and automatic weather stations. Since most automatic weather stations only have very short records (since 1990), we will not include these stations in our analysis. Table 3.1 shows the location and available records of the five meteorological stations in the Antarctic Peninsula. In our spatio-temporal analysis, we use the same period for the temperatures analysis at different locations. To keep the balance between longer records and more locations, we decide to use the temperatures at Bellingshausen, Marambio, Faraday/Vernadsky, Orcadas and Rothera over the period from January 1978 to December 1998 in our analysis (the observations from January 1999 to December 2000 will be used for the prediction purpose).

Suppose we have two locations A and B, with latitude (θ_1, θ_2) and longitude (ϕ_1, ϕ_2) respectively, then the great circle distance between A and B is given by:

$$3693.0 \arccos[\sin\theta_1 \sin\theta_2 + \cos\phi_1 \cos\phi_2 \cos(\phi_1 - \phi_2)], \tag{3.1}$$

where 3963.0 is the radius (in miles) of the spherical earth. Therefore, the pairwise distances of the above five stations are given in Table 3.2.

Table 3.1 The list of selected locations

No.	Meteorological station	Period	Latitude	Longitude
1	Bellingshausen	Jan 1969–Dec 2008	62.2S	58.9W
2	Marambio	Jan 1971–Dec 2003	64.2S	56.7W
3	Faraday/Vernadsky	Jan 1951–Dec 2006	65.4S	64.4W
4	Orcadas	Jan 1904–Dec 2000	60.7S	44.7W
5	Rothera	Jan 1978–Dec 2008	67.5S	68.0W

Table 3.2 Distances (in miles) between different sites at the Antarctic Peninsula

Site	Bellingshausen	Marambio	Faraday/Vernadsky	Orcadas	Rothera
Bellingshausen	0.000	155.6	281.3	498.9	459.8
Marambio	155.6	0.000	248.3	465.5	402.1
Faraday/Vernadsky	281.3	248.3	0.000	713.7	179.3
Orcadas	498.9	465.5	713.7	0.000	860.2
Rothera	459.8	402.1	179.3	860.2	0.000

3.3 Multivariate Regression Models with Space-Time ARMA Errors

3.3.1 The Model

Let Y_{it} denote the observation at time t at location i where $i = 1, \ldots, N$. Let $\underline{X}_t = (X_{1t}, \ldots, X_{rt})'$ be a deterministic vector of covariates at time point t $(t = 1, 2, \ldots, T)$. Let $\underline{Y}_t = (Y_{1t}, \ldots, Y_{Nt})'$ then $\underline{Y}_t$ follows multivariate regression models with space-time ARMA $(p_{\lambda_1,\ldots,\lambda_p}, q_{m_1,\ldots,m_q})$ errors if it can be written in the form:

$$\begin{cases} \underline{Y}_t = f(\underline{X}_t, \underline{\beta}) + \underline{Z}_t \\ \underline{Z}_t = \sum\limits_{k=1}^{p} \sum\limits_{l=0}^{\lambda_k} \phi_{kl} W^{(l)} \underline{Z}_{t-k} + \sum\limits_{k=1}^{q} \sum\limits_{l=0}^{m_k} \theta_{kl} W^{(l)} \underline{\varepsilon}_{t-k} + \underline{\varepsilon}_t \end{cases} \tag{3.2}$$

where $t = h + 1, \ldots, T$ and $h = \max(p, q)$.

Though the methodology to be described here can cope with nonlinear regression model, here we will restrict ourselves to the linear regression models. For convenience, we assume:

$$\underline{Y}_t = B\underline{X}_t + \underline{Z}_t$$

and we consider STARMA $(p_{\lambda_1,\ldots,\lambda_p}, q_{m_1,\ldots,m_q})$ models for $\underline{Z}_t$.

In model (3.2), p is the autoregressive order and q is the moving average order. λ_k is the spatial order of the kth autoregressive term and m_k is the spatial order of the kth moving average term. The scalars ϕ_{kl} and θ_{kl} are the autoregressive and moving average parameters at temporal lag k and spatial lag l respectively. $W^{(l)}$ is a non-zero $N \times N$ known weighting matrix at spatial lag l. Furthermore, $\{\underline{\varepsilon}_t\}$ is assumed to be i.i.d. random vectors, with:

$$E(\underline{\varepsilon}_t) = \underline{0}$$

and

$$E(\underline{\varepsilon}_t \underline{\varepsilon}'_{t+s}) = \begin{cases} \Sigma_\varepsilon, & s = 0 \\ 0, & s \neq 0. \end{cases}$$

It is further assumed that the zeros of

$$det\left(I_N + \sum_{k=1}^{p} \sum_{l=0}^{\lambda_k} \phi_{kl} W^{(l)} \underline{z}^k\right) \quad \text{and} \quad det\left(I_N + \sum_{k=1}^{p} \sum_{l=0}^{\lambda_k} \theta_{kl} W^{(l)} \underline{z}^k\right)$$

are outside the unit circle, which ensure the invertibility and the existence of a weakly stationary solution to the space-time ARMA models given by (3.2). We also assume that the above polynomials have no common zeros.

We have to define the weighting matrix based on the knowledge of the physical features of the spatial locations. Since the system under study is irregularly spaced, one reasonable approach to define the weighting matrices is to assume the spatial orders $\lambda_k = 1$ and $m_k = 1$ for all k and we have:

$$W^{(0)} = I_N \quad \text{and} \quad W^{(1)} = W$$

and the (i, j)th element of W is given by:

$$w_{ij} = \begin{cases} 1/d_{ij} & \text{i} \neq \text{j} \\ 0 & \text{i} = \text{j} \end{cases} \tag{3.3}$$

where d_{ij} is the distance between the sites i and j. The weights are also scaled such that $\sum_{j=1}^{N} w_{ij} = 1$, for each $i = 1, 2, \ldots, N$.

Such a definition of the weighting matrix is based on a reasonable idea: the temperatures from site i and j $(i, j = 1, \ldots, N)$ which are close to each other will be more correlated. Hence, a greater weight of the effect of temperatures at site j will be considered when we want to estimate the temperatures at site i.

By considering the particular weighting matrix described above and by restricting the function $f(\underline{X}_t, \underline{\beta})$ to be linear, the model we consider, as a special case of the model (3.2), is given by:

$$\begin{cases} \underline{Y}_t = B\underline{X}_t + \underline{Z}_t \\ \underline{Z}_t = \sum_{k=1}^{p} (\phi_{k0} I_N + \phi_{k1} W)\underline{Z}_{t-k} + \sum_{l=1}^{q} (\theta_{l0} I_N + \theta_{l1} W)\underline{\varepsilon}_{t-l} + \underline{\varepsilon}_t \end{cases} \tag{3.4}$$

where $t = h + 1, \ldots, T$ and $h = \max(p, q)$. B is a $N \times r$ matrix of unknown regression parameters. The $B_{m,n}$ $(m = 1, \ldots, N; n = 1, \ldots, r)$ element of the parameter matrix B represent the effect of the nth covariate on the mth location. The scalars ϕ_{k0}, ϕ_{k1} and θ_{k0}, θ_{k1} are the autoregressive and moving average parameters at temporal lag k and spatial lag 0 and 1. W is a non-zero $N \times N$ known weighting matrix which is defined in (3.3). $\{\underline{\varepsilon}_t\}$ is assumed to satisfy the conditions stated earlier. To ensure that there exists a weakly stationary solution and the invertibility is satisfied, similar conditions as model (3.2) are imposed.

3.3.2 Model Estimation and Model Selection

The model selection is done in two stages. Firstly, we choose the appropriate covariates in (3.4) under the assumption that $\{\underline{Z}_t\}$ are i.i.d. Then the residuals $\{\hat{\underline{Z}}_t\}$

are obtained by using least square estimation. We assume $\{\underline{\hat{Z}}_t\}$ follows a pure STARMA models and the autoregressive order p and the moving average order q are determined by minimizing BIC criterion given by:

$$T \ln |\hat{\Sigma}_\varepsilon| + 2k \ln(T) \tag{3.5}$$

where k is the number of parameters in the STARMA model. T is the number of observations and $\hat{\Sigma}_\varepsilon$ is the estimate of the covariance matrix of $\{\underline{\varepsilon}_t\}$ obtained by the methods of model estimation described later.

Suppose the autoregressive order p and the moving average order q have been identified, to estimate the regression parameters $B_{m,n}, (m = 1, \ldots, N; n = 1, \ldots, r)$, the autoregressive parameters ϕ_{k0}, ϕ_{k1} $(k = 1, \ldots, p)$ and the moving average parameters θ_{l0}, θ_{l1} $(l = 1, \ldots, q)$ in model (3.4), the weighted least square estimation procedure is used. The weighted sum of squares is defined by:

$$S = \sum_{t=h+1}^{T} \underline{\varepsilon}_t' \hat{\Sigma}_\varepsilon^{-1} \underline{\varepsilon}_t, \quad t = h+1, \ldots, T \tag{3.6}$$

where

$$\underline{\varepsilon}_t = \underline{Y}_t - B\underline{X}_t - \sum_{k=1}^{p} (\phi_{k0} I_N + \phi_{k1} W)(\underline{Y}_{t-k} - B\underline{X}_{t-k}) - \sum_{l=1}^{q} (\theta_{l0} I_N + \theta_{l1} W)\underline{\varepsilon}_{t-l}$$

and $\hat{\Sigma}_\varepsilon$ is the current estimate of the covariance matrix of $\underline{\varepsilon}_t$.

The weighted sum of squares can be minimized by using Newton–Raphson technique. This technique provides iterative estimates of the parameters (see [6]) but good initial estimates are still necessary. The initial estimate of regression parameter B is obtained by assuming that $\underline{Z}_t$ is a sequence of independent random variables. For the regression model with STAR errors, the initial estimates of autoregressive parameters are obtained by regressing $\underline{\hat{Z}}_t$ on $\underline{\hat{Z}}_{t-1}, W\underline{\hat{Z}}_{t-1}, \ldots, \underline{\hat{Z}}_{t-p}, W\underline{\hat{Z}}_{t-p}$. For the regression model with STARMA errors, we first generate $\{\hat{\varepsilon}_t\}$ through a regression model with high order STAR errors, then the initial estimates of autoregressive and moving average parameters are given by regressing $\underline{\hat{Z}}_t$ on $\underline{\hat{Z}}_{t-1}, W\underline{\hat{Z}}_{t-1}, \underline{\hat{\varepsilon}}_{t-1}, W\underline{\hat{\varepsilon}}_{t-1} \ldots$. The details of the procedure to find the initial estimates for STARMA models was described in [10].

When the fitted model is obtained, it is important to test whether the residuals $\{\underline{\hat{\varepsilon}}_t\}$ are independent. The sample space-time autocovariance function of $\underline{\hat{\varepsilon}}_t$ is defined by:

$$\hat{\gamma}_{lk}(s) = \text{trace}\left(\frac{[W^{(k)}]' W^{(l)} \hat{\Sigma}_\varepsilon(s)}{N}\right)$$

where

$$\hat{\Sigma}_\varepsilon(s) = \sum_{t=1}^{T-s} \frac{(\underline{\hat{\varepsilon}}_t - \underline{\bar{\hat{\varepsilon}}})(\underline{\hat{\varepsilon}}_{t+s} - \underline{\bar{\hat{\varepsilon}}})'}{T-s}.$$

The sample space-time autocorrelation function is given by:

$$\hat{\rho}_{lk}(s) = \frac{\widehat{\gamma}_{lk}(s)}{[\widehat{\gamma}_{ll}(0)\widehat{\gamma_{kk}}(0)]^{1/2}}.$$

Pfeifer and Deutsch [7] suggested that specific patterns can be observed by plotting $\hat{\rho}_{l0}(s)$,$(s = 1, 2, \ldots)$ for $l = 0, 1$, if residuals $\{\underline{\hat{\varepsilon}}_t\}$ are autocorrelated.

3.4 Asymptotic Properties and Simulations

Now we consider the model (3.4) described above. In this model, suppose $\underline{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{Nt})'$ and covariate vector $\underline{X}_t = (X_{1t}, \ldots, X_{rt})'$. Let w_{ij} denote the (i, j)th element of the weighting matrix W. The weighted sum of squares in (3.6) is denoted by S. To derive the asymptotic properties of the least square estimates of the model (3.4), we make the following assumptions:

- $\{\underline{\varepsilon}_t\}$ are i.i.d. with mean zero and variance $\sigma^2 I_N$.
- The zeros of

$$det\left(I_N + \sum_{k=1}^{p}(\phi_{k0} + \phi_{k1}W)\underline{z}^k\right) \quad \text{and} \quad det\left(I_N + \sum_{k=1}^{p}(\theta_{k0} + \theta_{k1}W)\underline{z}^k\right)$$

  are outside the unit circle.
- $E(\underline{Z}'_{t-k}\underline{Z}_{t-k'})$ is finite for $k, k' = 1, \ldots, p$.
- The covariates X_{it} are bounded, and further for fixed $i, j = 1, \ldots, r$ and $k, l = 1, \ldots, p$

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=h+1}^{T} X_{i(t-k)}X_{j(t-l)}$$

  exists and the $r \times r$ matrix with (i, j)th element given by $\lim_{T\to\infty}\frac{1}{T}\sum_{t=h+1}^{T} X_{it}X_{jt}$, is positive definite.

Furthermore, we assume $\Sigma_\varepsilon = \sigma^2 I_N$ for simplicity. Under this assumption, the sum of squares is given by:

$$S = \frac{1}{\sigma^2}\sum_{t=h+1}^{T}\underline{\hat{\varepsilon}}'_t\underline{\hat{\varepsilon}}_t, \quad t = h+1, \ldots, T.$$

Theorem 3.1. *Suppose* $\Psi = \{B_{m,n}, \phi_{k0}, \phi_{k1}, \theta_{l0}, \theta_{l1}\}$, *with* $m = 1, \ldots, N; n = 1, \ldots, r; k = 1, \ldots, p; l = 1, \ldots, q$. *The least square estimates* $\hat{\Psi}$ *are asymptotically normally distributed with mean* Ψ *and with covariance matrix given by:*

$$\frac{2}{T}\mathbf{C}^{-1} = \frac{2}{T}\begin{bmatrix} \mathbf{C_{11}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{22} & \mathbf{C}_{23} & \mathbf{C}_{24} & \mathbf{C}_{25} \\ \mathbf{0} & \mathbf{C}_{32} & \mathbf{C}_{33} & \mathbf{C}_{34} & \mathbf{C}_{35} \\ \mathbf{0} & \mathbf{C}_{42} & \mathbf{C}_{43} & \mathbf{C}_{44} & \mathbf{C}_{45} \\ \mathbf{0} & \mathbf{C}_{52} & \mathbf{C}_{53} & \mathbf{C}_{54} & \mathbf{C}_{55} \end{bmatrix}^{-1}$$

where

$$\mathbf{C}_{11} = \lim_{T\to\infty} \frac{1}{T} \frac{\partial^2 S}{\partial B_{m,n} \partial B_{m',n'}} = \frac{2}{\sigma^2} \lim_{T\to\infty} \frac{1}{T} \sum_{t=h+1}^{T} \sum_{i=1}^{N} \frac{\partial \varepsilon_{it}}{\partial B_{m,n}} \frac{\partial \varepsilon_{it}}{\partial B_{m',n'}}$$

with $m, m' = 1, \ldots, N; n, n' = 1, \ldots, r$, and

$$\mathbf{C}_{22} = E\left(\frac{1}{T} \frac{\partial^2 S}{\partial \phi_{k0} \partial \phi_{k'0}}\right) = \frac{1}{T\sigma^2} \sum_{t=h+1}^{T} 2E(\underline{Z}'_{t-k} \underline{Z}_{t-k'})$$

$$\mathbf{C}_{33} = E\left(\frac{1}{T} \frac{\partial^2 S}{\partial \phi_{k1} \partial \phi_{k'1}}\right) = \frac{1}{T\sigma^2} \sum_{t=h+1}^{T} 2E((W\underline{Z}_{t-k})' W \underline{Z}_{t-k'})$$

$$\mathbf{C}_{23} = \mathbf{C}_{32} = E\left(\frac{1}{T} \frac{\partial^2 S}{\partial \phi_{k0} \partial \phi_{k'1}}\right) = \frac{1}{T\sigma^2} \sum_{t=h+1}^{T} 2E((W\underline{Z}_{t-k})' \underline{Z}_{t-k'})$$

with $k, k' = 1, \ldots, p$, and

$$\mathbf{C}_{44} = E\left(\frac{1}{T} \frac{\partial^2 S}{\partial \theta_{l0} \partial \theta_{l'0}}\right) = \frac{1}{T\sigma^2} \sum_{t=h+1}^{T} 2E(\underline{\varepsilon}'_{t-l} \underline{\varepsilon}_{t-l'} + \underline{\varepsilon}'_{t-l-l'} \underline{\varepsilon}_t)$$

$$\mathbf{C}_{55} = E\left(\frac{1}{T} \frac{\partial^2 S}{\partial \theta_{l1} \partial \theta_{l'1}}\right) = \frac{1}{T\sigma^2} \sum_{t=h+1}^{T} 2E((W\underline{\varepsilon}_{t-l})' W \underline{\varepsilon}_{t-l'} + (W\underline{\varepsilon}_{t-l-l'})' W \underline{\varepsilon}_t)$$

$$\mathbf{C}_{45} = \mathbf{C}_{54} = E\left(\frac{1}{T} \frac{\partial^2 S}{\partial \theta_{l0} \partial \theta_{l'1}}\right) = \frac{1}{T\sigma^2} \sum_{t=h+1}^{T} 2E((W\underline{\varepsilon}_{t-l})' \underline{\varepsilon}_{t-l'} + (W\underline{\varepsilon}_{t-l-l'})' \underline{\varepsilon}_t)$$

with $l, l' = 1, \ldots, q$, and

$$\mathbf{C}_{24} = \mathbf{C}_{42} = E\left(\frac{1}{T}\frac{\partial^2 S}{\partial\phi_{k0}\partial\theta_{l0}}\right) = \frac{1}{T\sigma^2}\sum_{t=h+1}^{T} 2E(\underline{Z}'_{t-k}\underline{\varepsilon}_{t-l})$$

$$\mathbf{C}_{35} = \mathbf{C}_{53} = E\left(\frac{1}{T}\frac{\partial^2 S}{\partial\phi_{k1}\partial\theta_{l1}}\right) = \frac{1}{T\sigma^2}\sum_{t=h+1}^{T} 2E((W\underline{Z}_{t-k})'W\underline{\varepsilon}_{t-l})$$

$$\mathbf{C}_{25} = \mathbf{C}_{52} = E\left(\frac{1}{T}\frac{\partial^2 S}{\partial\phi_{k0}\partial\theta_{l1}}\right) = \frac{1}{T\sigma^2}\sum_{t=h+1}^{T} 2E((W\underline{Z}_{t-k})'\underline{\varepsilon}_{t-l})$$

$$\mathbf{C}_{34} = \mathbf{C}_{43} = E\left(\frac{1}{T}\frac{\partial^2 S}{\partial\phi_{k1}\partial\theta_{l0}}\right) = \frac{1}{T\sigma^2}\sum_{t=h+1}^{T} 2E((\underline{Z}_{t-k})'W\underline{\varepsilon}_{t-l}).$$

Proof. To prove Theorem 3.1, we follow the method of [8] given for univariate regression models with ARMA errors. We will first show that the first order derivatives of S with respect to Ψ evaluated at true values asymptotically converge to zero and the second and third derivatives evaluated at true values converge to some constants in probability. Then we verify that the estimates $\hat{\Psi}$ are consistent and evaluate the covariance matrix by a first order Taylor expansion of the equation:

$$\frac{1}{T}\frac{\partial S}{\partial\hat{\Psi}} = 0.$$

Finally we will establish the normality of the estimates.

As the first step to the proof for Theorem 3.1, we will show that:

$$\frac{1}{T}\frac{\partial S}{\partial\Psi} \to 0 \quad \text{in probability as} \quad T \to \infty$$

where $\Psi = \{B_{m,n}, \phi_{k0}, \phi_{k1}, \theta_{l0}, \theta_{l1}\}, m = 1, \ldots, N; n = 1, \ldots, r; k = 1, \ldots, p;$ $l = 1, \ldots, q$ and $h = \max(p, q)$.

Differentiating S with respect of $B_{m,n}$, we have:

$$\frac{\partial S}{\partial B_{m,n}} = \frac{2}{\sigma^2}\sum_{t=h+1}^{T}\sum_{i=1}^{N}\frac{\partial\varepsilon_{it}}{\partial B_{m,n}}\varepsilon_{it}$$

where

$$\frac{\partial \varepsilon_{it}}{\partial B_{m,n}} = \begin{cases} -X_{nt} + \sum_{k=1}^{p} (\phi_{k1} w_{ii} + \phi_{k0}) X_{n(t-k)} & \text{if i = m} \\ \sum_{k=1}^{p} (\phi_{k1} w_{ii} + \phi_{k0}) X_{n(t-k)} & \text{if i} \neq \text{m} \end{cases}$$

ε_{it} are independent with mean zero and $\frac{\partial \varepsilon_{it}}{\partial B_{m,n}}$ are finite and independent of $\underline{\varepsilon}_t$. Consequently, from the weak law of large numbers, we have:

$$\frac{1}{T} \frac{\partial S}{\partial B_{m,n}} \to 0 \quad \text{in probability as} \quad T \to \infty.$$

Differentiating S with respect of ϕ_{k0}, we have:

$$\frac{\partial S}{\partial \phi_{k0}} = \frac{2}{\sigma^2} \sum_{t=h+1}^{T} \sum_{i=1}^{N} \frac{\partial \varepsilon_{it}}{\partial \phi_{k0}} \varepsilon_{it}$$

where

$$\frac{\partial \varepsilon_{it}}{\partial \phi_{k0}} = -Y_{i(t-k)} + \sum_{n=1}^{r} B_{i,n} X_{n(t-k)} = -Z_{i(t-k)}, \quad \text{for all} \quad i = 1, \ldots, N.$$

Let $M_t = \frac{2}{\sigma^2} \sum_{i=1}^{N} Z_{i(t-k)} \varepsilon_{it}$, using the Markov's inequality, we have:

$$P\left(\left| \frac{1}{T} \sum_{t=h+1}^{T} M_t \right| \geq \varepsilon \right) \leq \frac{1}{\varepsilon^2} (T)^{-2} E \left(\sum_{t=h+1}^{T} M_t \right)^2, \quad \text{for all} \quad \varepsilon > 0.$$

It is easy to prove that $Z_{i(t-k)} \varepsilon_{it}$ have zero mean and uncorrelated over time t. Since M_t is sum of $Z_{t-k}(i) \varepsilon_{it}$, $E(M_t) = 0$ and $E(M_t M_s) = 0$ for $t \neq s$ and we have:

$$E \left(\sum_{t=h+1}^{T} M_t \right)^2 = \sum_{t=h+1}^{T} E(M_t)^2 + 2 \sum_{t>t'} E(M_t M_{t'}) = \sum_{t=h+1}^{T} E(M_t)^2,$$
$$\text{for} \quad t, t' = h+1, \ldots, T.$$

Furthermore, since $E(M_t)^2 < \infty$, we have:

$$P\left(\left| \frac{1}{T} \sum_{t=h+1}^{T} M_t \right| \geq \varepsilon \right) \quad \leq \quad \frac{1}{\varepsilon^2} (T)^{-2} \sum_{t=h+1}^{T} E(M_t)^2 \quad \leq \quad \frac{1}{\varepsilon^2} T^{-1} \max(E(M_t)^2).$$

Therefore:

$$\lim_{T\to\infty} P\left(\left|\frac{1}{T}\sum_{t=h+1}^{T} M_t\right| \geq \varepsilon\right) = 0 \quad \text{for all} \quad \varepsilon > 0$$

and hence:

$$\frac{1}{T}\sum_{t=h+1}^{T} M_t = \frac{1}{T}\frac{\partial S}{\partial \phi_{k0}} \to 0 \quad \text{in probability as} \quad T \to \infty, \quad \text{for} \quad k = 1, \ldots, p.$$

Differentiating S with respect of ϕ_{k1}, we have:

$$\frac{\partial S}{\partial \phi_{k1}} = \frac{2}{\sigma^2}\sum_{t=h+1}^{T}\sum_{i=1}^{N}\frac{\partial \varepsilon_{it}}{\partial \phi_{k1}}\varepsilon_{it}$$

where

$$\frac{\partial \varepsilon_{it}}{\partial \phi_{k1}} = -\sum_{u=1}^{N} w_{iu}(Y_{ut} - \sum_{n=1}^{r} B_{u,n} X_{n(t-k)}) = -\sum_{u=1}^{N} w_{iu} Z_{u(t-k)}.$$

Since W is a known matrix, the asymptotic properties of $\frac{1}{T}\frac{\partial S}{\partial \phi_{k1}}$ is similar to $\frac{1}{T}\frac{\partial S}{\partial \phi_{k0}}$. Therefore, the proof will be omitted.

Differentiating S with respect of θ_{l0}, we have:

$$\frac{\partial S}{\partial \theta_{l0}} = \frac{2}{\sigma^2}\sum_{t=h+1}^{T}\sum_{i=1}^{N}\frac{\partial \varepsilon_{it}}{\partial \theta_{l0}}\varepsilon_{it} \quad \text{where} \quad \frac{\partial \varepsilon_{it}}{\partial \theta_{l0}} = -\varepsilon_{i(t-l)}.$$

Since $\{\varepsilon_t\}$ is a sequence of zero mean i.i.d. random variables, we have $E(\varepsilon_{i(t-l)}\varepsilon_{it}) = 0$ for $l = 1, \ldots, q$.

We can also prove that $E(\varepsilon_{i(t-l)}\varepsilon_{it}, \varepsilon_{i(s-l)}\varepsilon_{is}) = 0$ for $t \neq s$. Following a similar procedure described above, we can show that:

$$\frac{1}{T}\frac{\partial S}{\partial \theta_{l0}} \to 0 \quad \text{in probability as} \quad T \to \infty \quad \text{for all} \quad l = 1, \ldots, q.$$

Similarly, since we have:

$$\frac{\partial S}{\partial \theta_{l1}} = \frac{2}{\sigma^2}\sum_{t=h+1}^{T}\sum_{i=1}^{N}\frac{\partial \varepsilon_{it}}{\partial \theta_{l1}}\varepsilon_{it} \quad \text{where} \quad \frac{\partial \varepsilon_{it}}{\partial \theta_{l1}} = -\sum_{u=1}^{N} w_{iu}\varepsilon_{u(t-l)}.$$

we can prove:

$$\frac{1}{T}\frac{\partial S}{\partial \theta_{l1}} \to 0 \quad \text{in probability as} \quad T \to \infty \quad \text{for all} \quad l = 1, \ldots, q.$$

Therefore, the first derivatives of S with respect to all the parameters $B_{m,n}, \phi_{k0}$, $\phi_{k1}, \theta_{l0}, \theta_{l1}$ converge to zero in probability.

Now we show:

$$\frac{1}{T}\frac{\partial^2 S}{\partial \Psi \partial \Psi'} \to \mathbf{C} \quad \text{in probability as} \quad T \to \infty$$

where $\Psi = \{B_{m,n}, \phi_{k0}, \phi_{k1}, \theta_{l0}, \theta_{l1}\}, m = 1, \ldots, N; n = 1, \ldots, r; k = 1, \ldots, p;$ $l = 1, \ldots, q$ and all the elements in the matrix have been defined earlier.

Obviously, we have:

$$\frac{1}{T}\frac{\partial^2 S}{\partial B_{m,n} \partial B_{m',n'}} \to \mathbf{C}_{11} \quad \text{in probability as} \quad T \to \infty.$$

For $\forall \varepsilon > 0$, by Markov's inequality, we have:

$$P\left(\mid \frac{1}{T}\frac{\partial^2 S}{\partial \phi_{k0} \partial \phi_{k'0}} - \mathbf{C}_{22} \mid > \varepsilon \right) \leq \frac{1}{\varepsilon} E\left(\mid \frac{1}{T}\frac{\partial^2 S}{\partial \phi_{k0} \partial \phi_{k'0}} - \mathbf{C}_{22} \mid \right).$$

Since

$$\mathbf{C}_{22} \quad \text{is defined as} \quad E\left(\frac{1}{T}\frac{\partial^2 S}{\partial \phi_{k0} \partial \phi_{k'0}} \right), \text{ we have:}$$

$$\lim_{T \to \infty} P\left(\left| \frac{1}{T}\frac{\partial^2 S}{\partial \phi_{k0} \partial \phi_{k'0}} - \mathbf{C}_{22} \right| > \varepsilon \right) = 0.$$

Therefore, we have:

$$\frac{1}{T}\frac{\partial^2 S}{\partial \phi_{k0} \partial \phi_{k'0}} \to \mathbf{C}_{22} \quad \text{in probability as} \quad T \to \infty.$$

Similarly, we can prove that the other terms converge to their corresponding expectations.

Differentiating S with respect to regression parameters $B_{m,n}$ and autoregressive parameters ϕ_{k0}, we have:

$$\frac{\partial^2 S}{\partial B_{m,n} \partial \phi_{k0}} = \frac{2}{\sigma^2} \sum_{t=h+1}^{T} \sum_{i=1}^{N} \left(\frac{\partial^2 \varepsilon_{it}}{\partial B_{m,n} \partial \phi_{k0}} \varepsilon_{it} + \frac{\partial \varepsilon_{it}}{\partial B_{m,n}} \frac{\partial \varepsilon_{it}}{\partial \phi_{k0}} \right)$$

where

$$\frac{\partial^2 \varepsilon_{it}}{\partial B_{m,n} \partial \phi_{k0}} = X_{n(t-k)} \quad \text{and} \quad \frac{\partial \varepsilon_{it}}{\partial \phi_{k0}} = Z_{i(t-k)}.$$

Since both $\frac{\partial^2 \varepsilon_{it}}{\partial B_{m,n} \partial \phi_{k0}}$ and $\frac{\partial \varepsilon_{it}}{\partial B_{m,n}}$ are finite, it is easy to show that:

$$\frac{1}{T} \frac{\partial^2 S}{\partial B_{m,n} \partial \phi_{k0}} \to 0 \quad \text{in probability as} \quad T \to \infty.$$

Now we consider:

$$\frac{\partial^2 S}{\partial B_{m,n} \partial \phi_{k1}} = \frac{2}{\sigma^2} \sum_{t=h+1}^{T} \sum_{i=1}^{N} \left(\frac{\partial^2 \varepsilon_{it}}{\partial B_{m,n} \partial \phi_{k1}} \varepsilon_{it} + \frac{\partial \varepsilon_{it}}{\partial B_{m,n}} \frac{\partial \varepsilon_{it}}{\partial \phi_{k1}} \right)$$

where

$$\frac{\partial^2 \varepsilon_{it}}{\partial B_{m,n} \partial \phi_{k1}} = w_{ii} X_{n(t-k)} \quad \text{and} \quad \frac{\partial \varepsilon_{it}}{\partial \phi_{k1}} = \sum_{u=1}^{N} w_{iu} Z_{u(t-k)}$$

and w_{ij} is the (i, j)th element of weighting matrix W for $(i, j = 1, \ldots, m)$. Obviously, the asymptotic properties for $\frac{\partial^2 \varepsilon_{it}}{\partial B_{m,n} \partial \phi_{k1}}$ is similar to $\frac{\partial^2 \varepsilon_{it}}{\partial B_{m,n} \partial \phi_{k0}}$, so we have:

$$\frac{1}{T} \frac{\partial^2 S}{\partial B_{m,n} \partial \phi_{k1}} \to 0 \quad \text{in probability as} \quad T \to \infty.$$

Since:

$$\frac{\partial^2 S}{\partial B_{m,n} \partial \theta_{l0}} = \frac{2}{\sigma^2} \sum_{t=h+1}^{T} \sum_{i=1}^{N} \left(\frac{\partial^2 \varepsilon_{it}}{\partial B_{m,n} \partial \theta_{l0}} \varepsilon_{it} + \frac{\partial \varepsilon_{it}}{\partial B_{m,n}} \frac{\partial \varepsilon_{it}}{\partial \theta_{l0}} \right)$$

where

$$\frac{\partial^2 \varepsilon_{it}}{\partial B_{m,n} \partial \theta_{l0}} = 0 \quad \text{and} \quad \frac{\partial \varepsilon_{it}}{\partial \theta_{l0}} = -\varepsilon_{i(t-l)}$$

by using the weak law of large numbers, we can easily show:

$$\frac{1}{T} \frac{\partial^2 S}{\partial B_{m,n} \partial \theta_{l0}} \to 0 \quad \text{and} \quad \frac{1}{T} \frac{\partial^2 S}{\partial B_{m,n} \partial \theta_{l1}} \to 0 \quad \text{in probability as} \quad T \to \infty.$$

Therefore, all the second order derivatives converges to matrix $\mathbf{C}$ which is given earlier.

Similarly, it is not difficult to prove that the third order derivatives of S are uniformly bounded, i.e., we have:

$$\frac{1}{T}\frac{\partial^3 S}{\partial\Psi_{i_1}\partial\Psi_{i_2}\partial\Psi_{i_3}} \to M_{i_1 i_2 i_3} \quad \text{in probability}$$

where $\Psi = \{B_{m,n}, \phi_{k0}, \phi_{k1}, \theta_{l0}, \theta_{l1}\}$, $m = 1,\ldots,N; n = 1,\ldots,r; k = 1,\ldots,p; l = 1,\ldots,q$ and $i_1, i_2, i_3 = 1,\ldots,mn + 2(p+q)$. $M_{i_1 i_2 i_3}$ is a constant independent of t.

Now we derive the covariance matrix of the estimators.

For convenience, let $u = mn + 2(p+q)$. Suppose Ψ_0 is the true parameter and $\hat{\Psi}$ is the least square estimate, also Ψ^* lie close to the true parameter. The Taylor expansion for the first order derivatives of S with respect to Ψ is:

$$\frac{1}{T}\frac{\partial S}{\partial\Psi}\mid_{\Psi=\hat{\Psi}} = \frac{1}{T}\frac{\partial S}{\partial\Psi}\mid_{\Psi=\Psi_0} + \frac{1}{T}(\hat{\Psi}-\Psi_0)\frac{\partial^2 S}{\partial\Psi\partial\Psi'}\mid_{\Psi=\Psi_0} + \frac{1}{2T}Q(\hat{\Psi}-\Psi_0)\mid_{\Psi=\Psi^*} \tag{3.7}$$

where the i_1th row of Q is given by:

$$(\hat{\Psi}-\Psi_0)'\frac{1}{T}\frac{\partial^3 S}{\partial\Psi_{i_1}\partial\Psi_{i_2}\partial\Psi_{i_3}}.$$

We have shown earlier that the first order derivatives on the right hand side of (3.7) converge to zero in probability and the second and third order derivatives converge to constants. Hence, by a theorem of [3], we have a consistent least square estimate of Ψ with probability approaching 1 as $T \to \infty$. Because the consistency of $\hat{\Psi}$, Q in (3.7) converges to zero. We also know that:

$$\frac{1}{T}\frac{\partial S}{\partial\Psi}\mid_{\Psi=\hat{\Psi}} = \mathbf{0} \quad \text{and} \quad \frac{1}{T}\frac{\partial^2 S}{\partial\Psi\partial\Psi'}\mid_{\Psi=\Psi_0} \to \mathbf{C} \quad \text{in probability as} \quad T\to\infty.$$

Hence, from (3.7), we have:

$$\frac{1}{\sqrt{T}}\frac{\partial S}{\partial\Psi}\mid_{\Psi=\Psi_0} + \sqrt{T}(\hat{\Psi}-\Psi_0)\mathbf{C} \to 0 \quad \text{as} \quad T\to\infty.$$

Therefore:

$$\frac{1}{\sqrt{T}}\mathbf{C}^{-1}\frac{\partial S}{\partial\Psi}\mid_{\Psi=\Psi_0} \quad \text{and} \quad \sqrt{T}(\hat{\Psi}-\Psi_0) \quad \text{have the same asymptotic distributions.}$$

Since we have shown earlier that the expectation of $\frac{1}{T}\frac{\partial S}{\partial\Psi}\mid_{\Psi=\Psi_0}$ equals zero, $(\hat{\Psi}-\Psi_0)$ will also have asymptotically zero mean. The sample covariance matrix of $\hat{\Psi}$ is

given by:

$$\Sigma = \frac{1}{T}\mathbf{C}^{-1} \lim_{T\to\infty} E\left(\frac{1}{T}\frac{\partial S}{\partial \Psi}\frac{\partial S}{\partial \Psi}\right)\mathbf{C}^{-1}.$$

It is straightforward to show that:

$$E\left(\frac{1}{T}\frac{\partial S}{\partial \Psi}\frac{\partial S}{\partial \Psi}\right) = 2E(\frac{1}{T}\frac{\partial^2 S}{\partial \Psi \partial \Psi'})$$

this implies:

$$\Sigma = \frac{1}{T}\mathbf{C}^{-1}2\mathbf{C}\mathbf{C}^{-1} = \frac{2}{T}\mathbf{C}^{-1}.$$

We now establish the asymptotic normality of the estimates. We will first study the asymptotic normality of the regression estimates $B_{mn}(m = 1,\ldots,N; n = 1,\ldots,r)$. We note that $\frac{\partial \varepsilon_{it}}{\partial B_{m,n}}$ are finite and independent of $\underline{\varepsilon}_t$. Therefore, $\frac{\partial S}{\partial B_{m,n}}$ is the sum of independent random variables with different variances. Since we have finite $\frac{\partial \varepsilon_{it}}{\partial B_{m,n}}$, it follows that:

$$\max_{1\le t\le T}\left(\frac{\partial \varepsilon_{it}}{\partial B_{m,n}}\right)^2 \Bigg/ \sum_{t=h+1}^{T}\sum_{i=1}^{N}\left(\frac{\partial \varepsilon_{it}}{\partial B_{m,n}}\right)^2 \to 0 \quad \text{as} \quad T\to\infty.$$

Hence, the Lindeberg condition is satisfied and $\frac{1}{\sqrt{T}}\frac{\partial S}{\partial B_{m,n}}$ are asymptotically normal by using the Lindeberg Feller central limit theorem. Since $\frac{1}{\sqrt{T}}I^{-1}\frac{\partial S}{\partial B_{m,n}}$ and $\sqrt{T}(\hat{B}_{m,n} - B_{(m,n)0})$ (where $B_{(m,n)0}$ is the true value of the regression parameter) have the same asymptotic distribution, it is straightforward to conclude that the estimate of regression parameters $B_{m,n}$ is also asymptotically normally distributed and any linear combination $\frac{1}{\sqrt{T}}\sum c_{m,n}\frac{\partial S}{\partial B_{m,n}}$ is also normally distributed. Therefore, $\frac{1}{\sqrt{T}}\frac{\partial S}{\partial B}$ is distributed asymptotically as a multivariate normal.

We now establish the asymptotic normality of the autoregressive and moving average estimates. From the previous proof, all the terms $\frac{\partial S}{\partial \phi_{k0}}, \frac{\partial S}{\partial \phi_{k1}}, \frac{\partial S}{\partial \theta_{l0}}, \frac{\partial S}{\partial \theta_{l1}}$ are linear sums of stationary and uncorrelated series. Moreover, they are linear sums of martingale differences. Hence, by a theorem of Billingsley [2], $\frac{1}{\sqrt{T}}\frac{\partial S}{\partial \phi_{k0}}, \frac{1}{\sqrt{T}}\frac{\partial S}{\partial \phi_{k1}}, \frac{1}{\sqrt{T}}\frac{\partial S}{\partial \theta_{l0}}, \frac{1}{\sqrt{T}}\frac{\partial S}{\partial \theta_{l1}}$ are asymptotically normal. Since any linear combination of above terms are asymptotically normal, the first order derivatives of S with respect to the autoregressive and moving average parameters are asymptotically multivariate normal.

Therefore, the least square estimators of all the parameters are asymptotically multivariate normally distributed and hence the results of Theorem 3.1. □

3.5 Simulation Study

The estimation methodology described above is used for simulated data. All the simulations were performed using pseudo-normal random numbers. Space-time data sets were simulated from different regression models with different STAR(1_1) and STARMA($1_1, 1_1$) errors. We consider a system of five locations with the weighting matrix given by:

$$W = \begin{bmatrix} 0.0000 & 0.4539 & 0.2510 & 0.1415 & 0.1536 \\ 0.4260 & 0.0000 & 0.2669 & 0.1424 & 0.1648 \\ 0.2441 & 0.2766 & 0.0000 & 0.0962 & 0.3831 \\ 0.2984 & 0.3199 & 0.2086 & 0.0000 & 0.1731 \\ 0.1907 & 0.2181 & 0.4892 & 0.1019 & 0.0000 \end{bmatrix}.$$

This weighting matrix is calculated based on the real locations of the five stations (Bellingshausen, Marambio, Faraday/Vernadsky, Orcadas and Rothera) in the Antarctic Peninsula.

3.5.1 Model 1

We have generated space-time data set from the model given by (the number of locations is assumed to be 5):

$$\begin{cases} \underline{Y}_t = B\underline{X}_t + \underline{Z}_t \\ \underline{Z}_t = (0.4I_N - 0.2W)\underline{Z}_{t-1} + \underline{\varepsilon}_t \end{cases} \quad t = 2, \ldots, T \quad \underline{\varepsilon}_t \sim MN(0, I_5) \tag{3.8}$$

where:

$$\underline{X}_t = \left(1, \frac{t}{T}, \cos\frac{2\pi t}{12}, \sin\frac{2\pi t}{12}\right)' \tag{3.9}$$

and

$$B = \begin{bmatrix} -5.0 & 2.0 & 3.0 & 3.0 \\ -5.0 & 1.5 & 3.0 & 3.0 \\ -5.0 & 2.0 & 3.0 & 3.0 \\ -5.0 & 1.5 & 3.0 & 3.0 \\ -5.0 & 2.0 & 3.0 & 3.0 \end{bmatrix}. \tag{3.10}$$

3.5.2 Model 2

We have also generated another time series from a regression model with STARMA errors which is given below:

$$\begin{cases} \underline{Y}_t = B\underline{X}_t + \underline{Z}_t \\ \underline{Z}_t = (0.4I_N - 0.2W)\underline{Z}_{t-1} \\ \qquad +(0.3I_N + 0.2W)\underline{\varepsilon}_{t-1} + \underline{\varepsilon}_t \end{cases} \quad t = 2, \ldots, T \quad \underline{\varepsilon}_t \sim MN(0, I_5) \quad (3.11)$$

where the covariates and the regression parameters B are same as in model 1.

3.5.3 Model 3

Besides, we have generated time series from the following regression model with STARMA errors:

$$\begin{cases} \underline{Y}_t = B\underline{X}_t + \underline{Z}_t \\ \underline{Z}_t = (0.3I_N + 0.1W)\underline{Z}_{t-1} \\ \qquad +(0.2I_N - 0.1W)\underline{\varepsilon}_{t-1} + \underline{\varepsilon}_t \end{cases} \quad t = 2, \ldots, T \quad \underline{\varepsilon}_t \sim MN(0, I_5) \quad (3.12)$$

where:

$$\underline{X}_t = \left(1, \frac{t}{T}, \cos\frac{2\pi t}{12}, \sin\frac{2\pi t}{12}\right)'$$

and

$$B = \begin{bmatrix} -5.0 & 2.5 & 3.0 & 3.0 \\ -5.0 & 2.5 & 3.0 & 3.0 \\ -5.0 & 2.5 & 3.0 & 3.0 \\ -5.0 & 2.5 & 3.0 & 3.0 \\ -5.0 & 2.5 & 3.0 & 3.0 \end{bmatrix}.$$

Tables 3.3, 3.4 and 3.5 show the average estimates of the parameters of interest from 100 replications for each T in the above three models. From the results of the tables we observe that the accuracy of the estimates increase with the sample size for all the models. Moreover, the standard deviations (given in the bracket of estimates) of all the estimates, as expected, decrease with the sample size. Therefore, the simulation results shows that the estimates are asymptotically consistent.

3.6 Real Data Analysis

In this section, we consider the analysis of the temperatures of Antarctic Peninsula by using multivariate regression models with space-time ARMA errors. The data described earlier consists of 252 monthly mean surface temperatures recorded

Table 3.3 Simulation results for model 1

Sample size	T = 100	T = 200	T = 400
$\hat{B}_{1,2}$	1.9998(0.1605)	2.0064(0.1224)	2.0047(0.0872)
$\hat{B}_{2,2}$	1.5067(0.1948)	1.4951(0.1157)	1.5061(0.0934)
$\hat{B}_{3,2}$	2.0317(0.1917)	1.9782(0.1200)	2.0039(0.0859)
$\hat{B}_{4,2}$	1.4873(0.1724)	1.4825(0.1170)	1.4977(0.0854)
$\hat{B}_{5,2}$	1.9980(0.1795)	2.0137(0.1150)	2.0122(0.0947)
$\hat{\phi}_{10}$	0.3969(0.0114)	0.3978(0.0084)	0.3994(0.0055)
$\hat{\phi}_{11}$	−0.2021(0.0207)	−0.1991(0.0150)	−0.2022(0.0113)

Table 3.4 Simulation results for model 2

Sample size	T = 100	T = 200	T = 400
$\hat{B}_{1,2}$	2.0159(0.2168)	2.0067(0.1472)	2.0044(0.1022)
$\hat{B}_{2,2}$	1.5161(0.2238)	1.5122(0.1801)	1.4990(0.1086)
$\hat{B}_{3,2}$	2.0042(0.2458)	2.0032(0.1552)	2.0067(0.1099)
$\hat{B}_{4,2}$	1.5133(0.2016)	1.5013(0.1535)	1.4910(0.1005)
$\hat{B}_{5,2}$	2.0233(0.2467)	1.9952(0.1604)	1.9822(0.1102)
$\hat{\phi}_{10}$	0.3963(0.0203)	0.3966(0.0144)	0.3974(0.0095)
$\hat{\phi}_{11}$	−0.2027(0.0370)	−0.2020(0.0293)	−0.1997(0.0183)
$\hat{\theta}_{10}$	0.3016(0.0223)	0.3028(0.0148)	0.3024(0.0091)
$\hat{\theta}_{11}$	0.2035(0.0363)	0.2014(0.0255)	0.2005(0.0157)

Table 3.5 Simulation results for model 3

Sample size	T = 100	T = 200	T = 400
$\hat{B}_{1,2}$	2.4904(0.1832)	2.4979(0.1093)	2.4938(0.0723)
$\hat{B}_{2,2}$	2.5080(0.1694)	2.5050(0.1363)	2.4947(0.0792)
$\hat{B}_{3,2}$	2.5066(0.1665)	2.4788(0.1240)	2.4965(0.0842)
$\hat{B}_{4,2}$	2.5143(0.1731)	2.4915(0.1187)	2.5049(0.0792)
$\hat{B}_{5,2}$	2.5015(0.1677)	2.5106(0.1440)	2.4969(0.0811)
$\hat{\phi}_{10}$	0.2961(0.0268)	0.2989(0.0207)	0.2974(0.0146)
$\hat{\phi}_{11}$	0.0995(0.0451)	0.0997(0.0330)	0.1062(0.0228)
$\hat{\theta}_{10}$	0.1990(0.0274)	0.1990(0.0209)	0.2020(0.0159)
$\hat{\theta}_{11}$	−0.0994(0.0489)	−0.0983(0.0365)	−0.1043(0.0253)

from January 1978 to December 1998 at the five meteorological stations (Bellingshausen, Marambio, Faraday/Vernadsky, Orcadas and Rothera) The time series plots of the temperatures for each station is given in Fig. 3.1. All the plots suggest strong seasonality of the series with period 12 months.

It is reasonable to use the inverse distance between two sites as the corresponding element of the weighting matrix since the correlation between the residuals $\{\hat{Z}_t\}$ after fitting the linear regression model in (3.13) from two sites seems to be inversely related to the distances (see Fig. 3.2). The corresponding stations of each correlation are shown in the label of each point and the number of each station is given in Table 3.1. For example, the point labeled $(2, 4)$ represents the distance

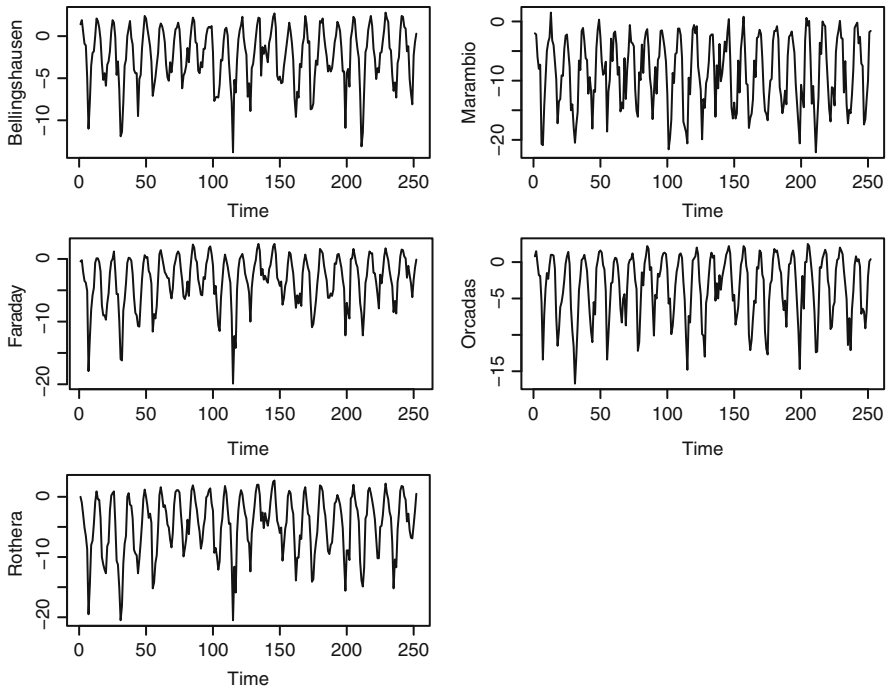


Fig. 3.1 Time series plots of Antarctic Peninsula

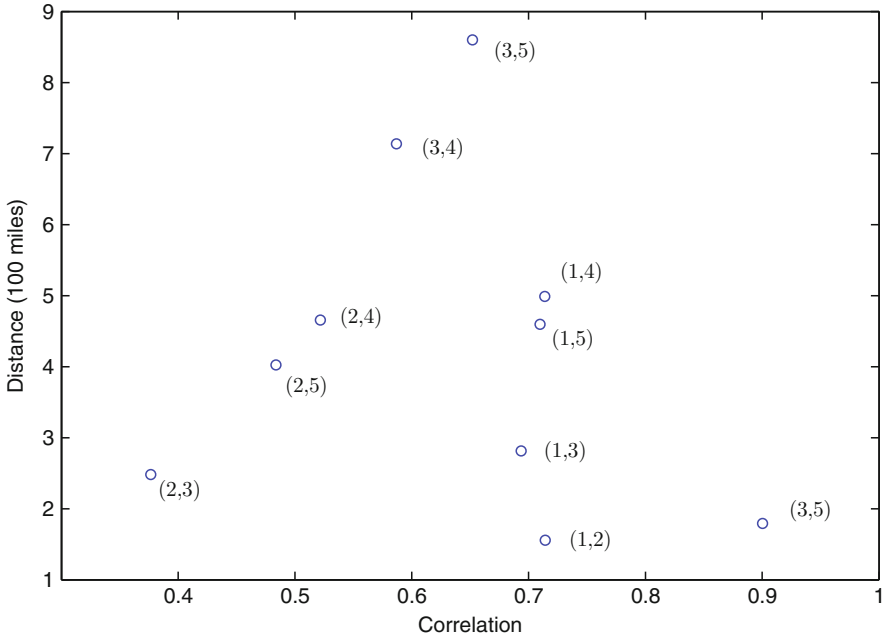


Fig. 3.2 Correlation between two series and the corresponding distance of the two sites

between Marambio and Orcadas and also the correlation between the residuals from the above two sites after fitting linear regression models.

The weighting matrix is calculated by the Definition 3.3, then it is standardized so that all rows add up to one, thus the weighting matrix W given in (3.8) is obtained. We use the following multivariate regression model with space-time ARMA errors considered earlier to fit the data at the five locations of Antarctic Peninsula:

$$\begin{cases} \underline{Y}_t = B\underline{X}_t + \underline{Z}_t \\ \underline{Z}_t = \sum_{k=1}^{p} (\phi_{k0} I_N + \phi_{k1} W)\underline{Z}_{t-k} + \sum_{l=1}^{q} (\theta_{l0} I_N + \theta_{l1} W)\underline{\varepsilon}_{t-l} + \underline{\varepsilon}_t \end{cases} \tag{3.13}$$

where $B_{5\times 4}$ is a matrix of parameters that need to be estimated. We have:

$$E(\underline{\varepsilon}_t) = \underline{0}$$

and

$$E(\underline{\varepsilon}_t \underline{\varepsilon}'_{t+s}) = \begin{cases} \Sigma_\varepsilon, & s = 0 \\ 0, & s \neq 0 \end{cases}$$

Since the temperature series of all the five locations have mean and trend and moreover they show strong seasonality with period 12, we use covariate vector $\underline{X}_t = [1, \frac{t}{T}, \cos\frac{2\pi t}{12}, \sin\frac{2\pi t}{12}]'$. Other assumptions of this model are same as (3.4). The autoregressive order p and the moving average order q are determined by minimizing BIC criterion. From the results in Table 3.6, we observe that the BIC is minimum when $p = 1$ and $q = 1$.

By using the estimation procedure described before, the estimates and corresponding standard errors (shown in the bracket) are calculated and are given in the following matrix:

$$\hat{B} = \begin{bmatrix} -2.4882(0.1723) & 0.3418(0.2925) & 3.1384(0.0829) & 2.6238(0.0835) \\ -8.5167(0.2403) & 0.0646(0.4089) & 5.6156(0.1267) & 2.2886(0.1276) \\ -4.5733(0.1807) & 2.2394(0.3070) & 3.0881(0.0896) & 3.5517(0.0900) \\ -3.6615(0.1875) & 0.5901(0.3186) & 4.2916(0.0943) & 3.1598(0.0949) \\ -5.7757(0.1978) & 2.2595(0.3362) & 4.4711(0.0993) & 4.0392(0.3169) \end{bmatrix}$$

Table 3.6 Model selection for STARMA models

Order	p = 0	1	2	3
q = 0		705.9	715.8	724.2
1	695.7	667.2	678.2	684.5
2	679.5	674.9	680.7	681.3
3	686.5	685.2	680.7	691.3

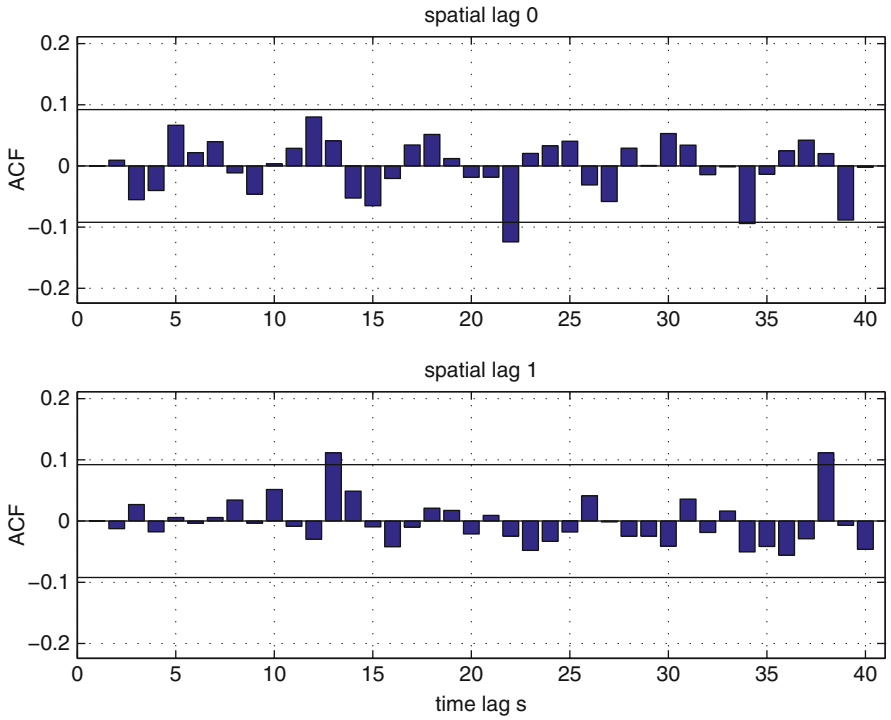


Fig. 3.3 Space time ACF for residuals $\{\underline{\hat{\varepsilon}}_t\}$

$$\hat{\phi}_{10} = 0.3374(0.0607) \quad \hat{\phi}_{11} = 0.2594(0.0727)$$
$$\hat{\theta}_{10} = -0.0119(0.0541) \quad \hat{\theta}_{11} = -0.1084(0.0559).$$

The results in the above table show that the estimates of mean (first column of $\hat{B}$) and harmonic components (third and fourth column of $\hat{B}$) at all the locations are significant. The trends (second column of $\hat{B}$) are significant at level 1% at Faraday/Vernadsky and Rothera. The trends are also significant at level 10% at Orcadas but they are not significant at other locations. This analysis confirms that the warming at Faraday/Vernadsky and Rothera are very significant, compared to other stations. The results also show that the estimates of autoregressive parameters and the moving average parameter at spatial lag 1 ($\hat{\theta}_{11}$) are significant.

Figure 3.3 shows that the residuals $\{\underline{\hat{\varepsilon}}_t\}$ obtained from the fitted model are not autocorrelated since autocorrelations at most lags are not significant and there are reasons that they are normal and hence the assumption that the residuals are independent is satisfied.

After fitting the model to our temperature data, we are interested in forecasting the temperatures of each site in the future. Let $\underline{Y}_t$ be an vector of observations at each site at time t. Let $\widetilde{\underline{Y}}_t(l)$ be the predictor of $\underline{Y}_{t+l}$ made at time t. The mean square error of the l step predictor is given by:

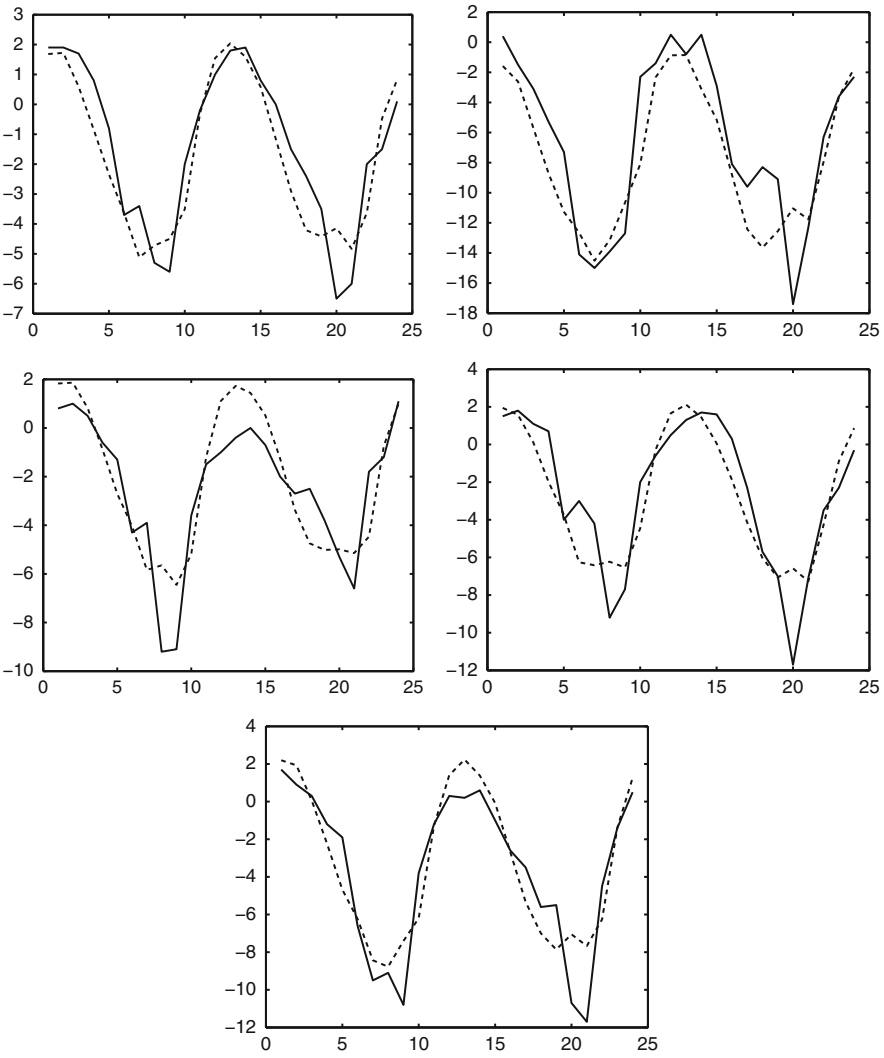


Fig. 3.4 Predictions of the temperatures from January 1999 to December 2000

$$M(l) = E[(\widetilde{\underline{Y}}_t(l) - \underline{Y}_{t+l})^2].$$

The optimum predictor of $\underline{Y}_{t+l}$ which minimizes the mean square error is:

$$\widetilde{\underline{Y}}_{t+l} = E(\underline{Y}_{t+l} | \underline{Y}_t, \underline{Y}_{t-1}, \ldots).$$

For our model, the one step predictor for the temperatures is given by:

$$\widetilde{\underline{Y}}_{t+1} = \hat{B} X_{t+1} + (\hat{\phi}_{10} I_5 + \hat{\phi}_{11} W)(\underline{Y}_t - \hat{B}\underline{X}_t) + (\hat{\theta}_{10} I_5 - \hat{\theta_{11}} W)\hat{\underline{\varepsilon}}_t.$$

The one step prediction for the temperatures from January 1999 to December 2000 at the five stations are compared with the real data over the same period. The comparisons of the predictions with the true values are given in Fig. 3.4 (the dash line is the prediction line). The results show that all the predictions are close to the true observations.

3.7 Conclusion

In this paper we examined the temporal and spatial variability of the temperatures at the five locations in the Antarctic Peninsula. Besides fitting linear regression models to temperatures, we consider multivariate regression models with space-time ARMA errors to study the spatio-temporal variation. From our analysis, the monthly mean surface temperatures at all the locations in the Antarctic Peninsula are increasing over the period from January 1978 to December 1998. The largest warming trends are found at the Faraday/Vernadsky and Rothera, with the warming rate of 1.0664 °C/decade and 1.0759 °C/decade respectively. The trends at the above two locations are also significant at 1% level. The temperatures at Orcadas is also increasing at 10% significance level, with the warming rate of 0.5901 °C/decade.

A warming rate of 1.01 °C/decade at Rothera over the period from 1978 to 2000 is also reported by [12]. However, the estimated trend given in [12] is not significant because of the lack of efficiency by using simple linear regression models to fit temperature data. From our analysis, we believe the substantial warming occurred at Rothera station is significant at level 1% since the seasonality and temporal autocorrelations have been taken into account.

References

1. Antunes AM, Subba Rao T (2006) On hypotheses testing for the selection of spatio-temporal models. J Time Series Anal 27:767–791
2. Billingsley P (1961) The Lindeberg–Lévy theorem for martingales. Proc Am Math Soc 12:788–792
3. Chanda KC (1954) A note on the consistency and maxima of the roots of the likelihood equations. Biometrika 41:56–61
4. Folland C, Karl T, Christy J, Clarke R, Gruza G, Jouzel J, Mann ME, Oerlemans J, Salinger M, Wang S-W, et al (2001) Observed climate variability and change. In: Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, Dai X, Maskell K, Johnson CA (eds) Climate change 2001: The scientific basis. Cambridge University Press, Cambridge, pp 99–182
5. King JC, Comiso JC (2003) The spatial coherence of interannual temperature variations in the Antarctic Peninsula. Geophys Res Lett 30(2):1040. doi:10.1029/2002GL015580
6. Pfeifer PE, Deutsch SJ (1980a) A three-stage iterative procedure for space-time modelling. Technometrics 22(1):35–47
7. Pfeifer PE, Deutsch SJ (1980b) Identification and interpretation of first order space-time ARMA models. Technometrics 22(3):397–408
8. Pierce DA (1971) Least square estimation in the regression model with autoregressive-moving average errors. Biometrika 58(2):299–312

9. Santer BD, Wiggley TML, Boyle JS, Gaffen DJ, Hnilo JJ, Nychka D, Parker DE, Taylor KE (2000) Statistical significance of trends and trend differences in layer-average atmospheric temperature time series. J Geophys Res 105:7227–7356
10. Subba Rao T, Antunes AM (2003) Spatio-temporal modelling of temperature time series: a comprative study. In: Schonberg F, Brillinger DR, Robinson E (eds) Time series analysis and applications to geophysical systems, IMA Publications, Springer, New York, 123–150
11. Trenberth KE (1984) Signal versus noise in the southern oscillation. Mon Weather Rev 112:326–332
12. Turner J, Colwell SR, Marshall GJ, Lachlan-Cope TA, Carleton AM, Jones PD, Lagun V, Reid PA, Iagovkina S (2005) Antarctic climate changes during the last 50 years. Int J Climatol 25:279–294
13. Vaughan DG, Marshall GJ, Connolley WM, Parkinson C, Mulvaney R, Hodgson DA, King JK, Pudsey CJ, Turner J (2003) Recent rapid regional climate warming on the Antarctic Peninsula. Clim Change 60:243–274

Chapter 4
The Generalized von Mises–Fisher Distribution

Riccardo Gatto

Abstract In this chapter we introduce the broad class of generalized von Mises–Fisher (GvMF) distributions on the unit hypersphere S^{p-1}, which arises from a generalization of the von Mises–Fisher distribution. This class of distributions has some important information theoretic properties. It is shown that, under constraints on some moments along some fixed directions, and using Kullback–Leibler information as measure, the closest distribution to any predetermined one on S^{p-1}, has the GvMF form. Lower bounds for the Kullback–Leibler information in this context are also provided. Several connections between GvMF and other directional or linear distributions are given. GvMF distributions can be re-expressed in terms of generalized von Mises distributions when $p = 2$. GvMF distributions arise as offset multivariate normal distributions, as offset singular normal distributions and as offset distributions from an exponential spherical type of distributions. Various forms of GvMF densities which feature uni- and multimodality, with different shape of modes, girdle and with other particularities are graphically illustrated.

4.1 Introduction

In this chapter we introduce the generalized von Mises–Fisher (GvMF) directional probability distribution, motivated from its high flexibility, its information theoretic properties and its characterizations in terms of other important distributions. Directional data which are directions in R^p, $p = 2, 3, \ldots$, and can be represented by points on the surface of the $(p-1)$-dimensional unit hypersphere $S^{p-1} = \{\underline{x} \in \mathrm{R}^p | \underline{x}^{\mathrm{T}}\underline{x} = 1\}$. Most practical situations are for circular data, where $p = 2$, and for spherical data, where $p = 3$. Many results of this chapter are given for the general case of $p > 3$, because it is conceptually similar to the case of $p = 3$. Spherical data arise in various fields such as physics, astronomy and earth sciences

R. Gatto
Department of Mathematics and Statistics, University of Bern, 3012 Bern, Switzerland,
e-mail: gatto@stat.unibe.ch

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_4,

and can have experimental or observational origins. A probability distribution on S^{p-1} is a directional probability distribution. Two general references on directional statistics are [10] and [6]. In the following we are only concerned with absolutely continuous directional distributions with respect to (w.r.t.) the uniform distribution on S^{p-1}. When $p = 2$ it is generally convenient to represent a random unit vector by its angle w.r.t. a chosen null direction and with the radian measure, for example. This random angle has a circular density, w.r.t. the Lebesgue measure on S^1, which is a non-negative 2π-periodic function on R integrating to one on $[0, 2\pi)$. In general we can re-express $\underline{x} \in \mathrm{R}^p \backslash \{0\}$ in angular coordinates by the diffeomorphism $\underline{x} = \underline{g}(r, \theta_1, \ldots, \theta_{p-1})$ which, for $p \geq 3$, is given by

$$\begin{aligned} x_1 &= r \sin\theta_1 \sin\theta_2 \ldots \sin\theta_{p-2} \cos\theta_{p-1}, \\ x_2 &= r \sin\theta_1 \sin\theta_2 \ldots \sin\theta_{p-2} \sin\theta_{p-1}, \\ &\vdots \quad \vdots \\ x_{p-1} &= r \sin\theta_1 \sin\theta_2, \\ x_p &= r \cos\theta_1, \end{aligned}$$

where $r > 0, 0 \leq \theta_i \leq \pi$ for $i = 1, \ldots, p-2$ and $0 \leq \theta_{p-1} < 2\pi$. For $p = 3$, θ_1 is the co-latitude and θ_2 is the longitude. The determinant of the Jacobian of $\underline{g}$ is given by $J(r; \theta_1, \ldots, \theta_{p-2}) = r^{p-1} \prod_{j=2}^{p-1} \sin^{p-j} \theta_{j-1}$, for $p \geq 3$, and it is equal to r for $p = 2$. Jacobi's theorem yields that for any continuous function $f : S^{p-1} \to \mathrm{R}$,

$$\int_{S^{p-1}} f(\underline{x}) d\sigma(\underline{x}) = \int_0^{2\pi} \int_0^{\pi} \ldots \int_0^{2\pi} f(\underline{g}(1; \theta_1, \ldots, \theta_{p-1})) J(1; \theta_1, \ldots, \theta_{p-2}) d\theta_1 \ldots d\theta_{p-1},$$

where $d\sigma(\underline{x})$ denotes the infinitesimal surface area on S^{p-1} and around $\underline{x}$ on S^{p-1}. From this, the surface area of S^{p-1} is $a_p = \int_{S^{p-1}} d\sigma(\underline{x}) = 2\pi^{p/2}/\Gamma(p/2)$, Γ denoting the gamma function. Hence, the uniform or isotropic probability distribution over S^{p-1} is $U[A] = \int_A d\sigma(\underline{x})/a_p$, for $A \subset S^{p-1}$, which is the unique invariant distribution under rotation and reflection. For a probability distribution P over S^{p-1} and for $A \subset S^{p-1}$, $P[A] = \int_A (dP)/(dU) dU$, where $(dP)/(dU)$ is the Radon–Nikodym derivative or density of P w.r.t. U. All directional densities presented in this chapter are given w.r.t. U.

We now introduce a class of probability distributions on S^{p-1} which generalizes and unifies several other existing distributions. The generalized von Mises-Fisher distribution of order k (GvMF_k) is the probability distribution P on S^{p-1} having the density w.r.t. the isotropic distribution U at $\underline{x} \in S^{p-1}$

$$f(\underline{x} | \underline{m}_1, \ldots, \underline{m}_k; c_1, \ldots, c_k; i_1, \ldots, i_k) \stackrel{\text{def}}{=} \frac{dP}{dU}(\underline{x}) \propto \exp\left\{ \sum_{r=1}^{k} c_r (\underline{m}_r^{\mathrm{T}} \underline{x})^{i_r} \right\}, \tag{4.1}$$

where $\underline{m}_r \in S^{p-1}$, $c_r \in \mathrm{R}$, for $r = 1, \ldots, k$, $i_1 \leq \ldots \leq i_k \in \{1, 2, \ldots\}$ and, to avoid over indentification, $1, (\underline{m}_1^{\mathrm{T}}\underline{x})^{i_1}, \ldots, (\underline{m}_k^{\mathrm{T}}\underline{x})^{i_k}$ are linearly independent functions of $\underline{x} \in S^{p-1}$. We denote as $\mathrm{GvMF}_k(\underline{m}_1, \ldots, \underline{m}_k; c_1, \ldots, c_k; i_1, \ldots, i_k)$ a representative random direction of R^p with density (4.1). The particular case of $k = 1$ and $i_1 = 1$ in (4.1) yields the von Mises–Fisher (vMF) density which plays an important role in directional statistics (see, e.g., [10, p. 168]), and we denote as $\mathrm{vMF}(\underline{m}_1; c_1)$, $\underline{m}_1 \in \mathrm{R}^p$, $c_1 \geq 0$, a representative random direction of R^p with this density. Thus, a GvMF_k with $i_1, \ldots, i_k$ known belongs to the most general exponential directional distribution of [2], with density proportional to the exponential of polynomials in $\underline{x} \in S^{p-1}$. The reparametrization (4.1) leads however to intuitive interpretations of the parameters, to information theoretic optimalities and to various characterization properties in terms of other important distributions.

In Sect. 4.2, we show that GvMF distributions possess some important information theoretic properties. We first show that for some fixed moments along some predetermined directions, the closest density, in the Kullback–Leibler sense, to a given density f_0, all of them on S^{p-1}, has the GvMF form (4.1). We also provide a lower bound for the Kullback–Leibler information under these moment restrictions as well as the entropy of GvMF distributions. In Sect. 4.3, we show that GvMF distributions are closely related with other important directional and linear distributions. When $p = 2$, any GvMF distribution admits a generalized von Mises (GvM) reparametrization. The GvM distribution was introduced by [5]. GvMF_2 distributions are closely related to the Fisher–Watson and the Fisher–Bingham directional distributions. GvMF distributions arise also as conditional offset distributions from the multivariate normal, the singular normal and from a type of spherical exponential distributions. Still in Sect. 4.3, we show various graphs of GvMF densities, which feature uni- and multimodality, different modal shapes, girdle and other interesting forms.

4.2 Information Theoretic Results for Directional Distributions

In this section we derive some optimal directional distributions w.r.t. directional versions of Kullback–Leibler information and entropy, respectively defined by (4.2) and (4.3) below. Under both criteria, either GvMF distributions or other directional distributions having a GvMF part turn out to be the optima. An important consequence is that these distributions are optimal solutions of a constrained prior selection problem of Bayesian statistics. In our setting, an optimal directional prior distribution is obtained by maximizing the entropy given some prior knowledge on the moments along some directions.

The Kullback–Leibler information was introduced by [8] and plays a central role in information theory. We suggest defining this quantity on S^{p-1} also. Suppose P and Q are two probability measures on S^{p-1} w.r.t. the uniform distribution U and that P is absolutely continuous w.r.t. Q, which is denoted $P \ll Q$. Then we define

$$I(P|Q) = \int_{S^{p-1}} \log \frac{dP}{dQ} dP = \int_{S^{p-1}} \log \frac{f(\underline{x})}{g(\underline{x})} f(\underline{x}) dU(\underline{x}), \tag{4.2}$$

where f and g are the directional densities of P and Q respectively w.r.t. U and where $0 \log 0 = 0$ by assumption. $I(P|Q)$ is the mean logarithmic likelihood ratio w.r.t. P or mean information per observation of P for discriminating of Q from P. The Gibbs inequality implies that $I(P|Q)$ is positive semi-definite, i.e.,

$$I(P|Q) \geq 0, \tag{4.3}$$

for all P and Q such that $P \ll Q$, with equality iff $P = Q$ a.s. $I(P|Q)$ is sometimes referred to as relative entropy or Kullback–Leibler distance, even though $I(P|Q)$ is not a metric: it violates the symmetry and the triangle rules. A related quantity is the entropy of [12], which we suggest extending to S^{p-1} by

$$H(P) = -\int_{S^{p-1}} \log \frac{dP}{dU} dP = -\int_{S^{p-1}} \log f(\underline{x}) f(\underline{x}) dU(\underline{x}),$$

where $0 \log 0 = 0$ is assumed. $H(P)$ measures the uncertainty inherent in P or in f. Equivalently, $H(P)$ measures the expected amount of information gained on obtaining a direction from P, based on the principle that the rarer an event, the more informative its occurrence. We denote alternatively $I(f|g)$ for $I(P|Q)$ and $H(f)$ for $H(P)$.

We denote by DM_r the rth directional moment condition on the density f given by

$$\int_{S^{p-1}} (\underline{m}_r^{\mathrm{T}} \underline{x})^{i_r} f(\underline{x}) dU(\underline{x}) = \alpha_r,$$

where $i_1 \leq \ldots \leq i_k \in \{1, 2, \ldots\}$, $\underline{m}_r \in S^{p-1}$, $\alpha_r \in \mathrm{R}$ and $r = 1, \ldots, k$.

All densities in the next theorem and corollaries are on S^{p-1} and w.r.t. the isotropic distribution U on S^{p-1}.

Theorem 4.1. *(a) The directional density f which satisfies DM_r, for $r = 1, \ldots, k$, and which minimizes the Kullback–Leibler information $I(f|f_0)$, w.r.t. a given directional density f_0, is proportional to*

$$\exp\left\{\sum_{r=1}^{k} c_r (\underline{m}_r^{T} \underline{x})^{i_r}\right\} f_0(\underline{x}), \tag{4.4}$$

for $\underline{x} \in S^{p-1}$. The parameters $c_1, \ldots, c_k$ are the solutions in $v_1, \ldots, v_k$ of

$$\frac{\partial}{\partial v_r} K(v_1, \ldots, v_k; \underline{m}_1, \ldots, \underline{m}_k; i_1, \ldots, i_k; f_0) = \alpha_r, \quad r = 1, \ldots, k, \tag{4.5}$$

where $K = \log M$ and

$$M(v_1,\ldots,v_k;\underline{m}_1,\ldots,\underline{m}_k;i_1,\ldots,i_k;f_0)=\int_{S^{p-1}}\exp\left\{\sum_{r=1}^{k}v_r(\underline{m}_r^T\underline{x})^{i_r}\right\}f_0(\underline{x})dU(\underline{x}). \tag{4.6}$$

(b) For any directional density f satisfying DM_r*, for $r=1,\ldots,k$, we have the following lower bound on the Kullback–Leibler information,*

$$I(f\,|\,f_0)\geq -K(c_1,\ldots,c_k;\underline{m}_1,\ldots,\underline{m}_k;i_1,\ldots,i_k;f_0)+\sum_{j=1}^{k}c_j\alpha_j, \tag{4.7}$$

with equality iff f is proportional to (4.4).

Proof. The basic idea of this proof is due to [7].
(a) Define $t_r(\underline{x})=(\underline{m}_r^{\mathrm{T}}\underline{x})^{i_r}$, $r=1,\ldots,k$. In order to minimize $I(f\,|\,f_0)$ w.r.t. f and under DM_r, for $r=1,\ldots,k$, we minimize w.r.t. f the Lagrangian

$$\begin{aligned}L(f)&=\int_{S^{p-1}}\left\{\log\frac{f(\underline{x})}{f_0(\underline{x})}-\sum_{r=1}^{k}v_rt_r(\underline{x})-v_0\right\}f(\underline{x})dU(\underline{x})\\&=\int_{S^{p-1}}\psi(q(\underline{x}))f_0(\underline{x})dU(\underline{x}),\end{aligned}$$

where

$$\psi(q)=\left(\log q-\sum_{r=1}^{k}v_rt_r-v_0\right)q$$

and $q=f/f_0$. Define

$$q_0=\exp\left\{\sum_{r=1}^{k}v_rt_r+v_0-1\right\}.$$

We note that $\psi(q_0)=-q_0$, $\psi'(q)=\log q-\sum_{r=1}^{k}v_rt_r-v_0+1$, $\psi'(q_0)=0$, $\psi''(q)=1/q$. For any function h on S^{p-1} between q and q_0, we have

$$\begin{aligned}L(f)&=\int_{S^{p-1}}\Big[\psi(q_0(\underline{x}))+\{q(\underline{x})-q_0(\underline{x})\}\psi'(q_0(\underline{x}))\\&\quad+\frac{1}{2}\{q(\underline{x})-q_0(\underline{x})\}^2\psi''(h(\underline{x}))\Big]f_0(\underline{x})dU(\underline{x})\\&=\int_{S^{p-1}}-q_0(\underline{x})f_0(\underline{x})dU(\underline{x})+\frac{1}{2}\int_{S^{p-1}}\{q(\underline{x})-q_0(\underline{x})\}^2h^{-1}(\underline{x})f_0(\underline{x})dU(\underline{x}).\end{aligned} \tag{4.8}$$

This implies

$$L(f) \geq -\int_{S^{p-1}} \exp\left\{\sum_{r=1}^{k} v_r t_r(\underline{x}) + v_0 - 1\right\} f_0(\underline{x}) dU(\underline{x}), \tag{4.9}$$

with equality iff the second integral in (4.8) vanishes, which happens iff

$$f = \exp\left\{\sum_{r=1}^{k} v_r t_r + v_0 - 1\right\} f_0.$$

From the definition (4.6) and because f integrates to one w.r.t. U on S^{p-1}, we have $e^{v_0-1} = M^{-1}(v_1, \ldots, v_k; \underline{m}_1, \ldots, \underline{m}_k; i_1, \ldots, i_k; f_0)$. Hence we obtain

$$f(\underline{x}) = M^{-1}(v_1, \ldots, v_k; \underline{m}_1, \ldots, \underline{m}_k; i_1, \ldots, i_k; f_0) \exp\left\{\sum_{r=1}^{k} v_r t_r(\underline{x})\right\} f_0(\underline{x}). \tag{4.10}$$

From DM_r, $r = 1, \ldots, k$, follows that $c_1, \ldots, c_k$ are the solutions in $v_1, \ldots, v_k$ of (4.5).
(b) From (4.8) and (4.9), we obtain that for any density f satisfying DM_r, $r = 1, \ldots, k$, (4.7) must hold and it must hold with an equality iff we take the optimal f given by (4.10). □

Corollary 4.1. *(a) For any $k \in \{1, 2, \ldots\}$, the directional density f maximizing the entropy $H(f)$ under DM_r, $r = 1, \ldots, k$, is the $\mathrm{GvMF}_k(\underline{m}_1, \ldots, \underline{m}_k; c_1, \ldots, c_k; i_1, \ldots, i_k)$ density. The parameters $c_1, \ldots, c_k$ are the solutions in $v_1, \ldots, v_k$ of*

$$\frac{\partial}{\partial v_r} K(v_1, \ldots, v_k; \underline{m}_1, \ldots, \underline{m}_k; i_1, \ldots, i_k; 1) = \alpha_r, \tag{4.11}$$

$r = 1, \ldots, k$.
(b) If f is a directional density satisfying DM_r, $r = 1, \ldots, k$, then

$$H(f) \leq K(c_1, \ldots, c_k; \underline{m}_1, \ldots, \underline{m}_k; i_1, \ldots, i_k; 1) - \sum_{r=1}^{k} c_r \alpha_r,$$

with equality iff f is the $\mathrm{GvMF}_k(\underline{m}_1, \ldots, \underline{m}_k; c_1, \ldots, c_k; i_1, \ldots, i_k)$ density.

Proof. (a) By choosing $f_0 = 1$ a.s. on S^{p-1}, we find $I(f|f_0) = -H(f)$. Hence maximizing $H(f)$ w.r.t. f under DM_r, $r = 1, \ldots, k$, is equivalent to minimizing $I(f|f_0)$ w.r.t. f under these constraints. It follows from Theorem 4.1(a) that this minimum is attained when f is a GvMF_k density. The parameters $c_1, \ldots, c_k$ which satisfy the constraints are given by (4.11), which is exactly (4.5) with $f_0 = 1$.
(b) This part follows directly from Theorem 4.1(b). □

As already mentioned, Corollary 4.1 can be used for prior selection in Bayesian statistics. An appealing property of the GvMF class of distributions is its closure w.r.t. directional moment restrictions in the following sense. All densities are again given w.r.t. the isotropic measure U on S^{p-1}. The proof of the following corollary is direct and omitted. In the next corollary, we relax the inessential restriction on the ordering $i_1 \leq \ldots \leq i_j \leq i_{j+1} \leq \ldots \leq i_{j+k}$.

Corollary 4.2. *(a) Denote by* f_0 *the* $\mathrm{GvMF}_j(\underline{m}_1, \ldots, \underline{m}_j; c_1, \ldots, c_j; i_1, \ldots, i_j)$ *density. Then the closest density* f *to* f_0*, in the sense of minimizing* $I(f \,|\, f_0)$*, which satisfies* DM_r*,* $r = j+1, \ldots, j+k$*, is the* $\mathrm{GvMF}_{j+k}(\underline{m}_1, \ldots, \underline{m}_{j+k}; c_1, \ldots, c_{j+k}; i_1, \ldots, i_{j+k})$ *density. The parameters* $c_{j+1}, \ldots, c_{j+k}$ *are solutions in* $v_{j+1}, \ldots, v_{j+k}$ *of*

$$\frac{\partial}{\partial v_r} K(v_{j+1}, \ldots, v_{j+k}; \underline{m}_{j+1}, \ldots, \underline{m}_{j+k}; i_{j+1}, \ldots, i_{j+k}; f_0) = \alpha_r, \quad r = j+1, \ldots, j+k.$$

(b) For any density f *satisfying* DM_r*,* $r = j+1, \ldots, j+k$*,*

$$I(f \,|\, f_0) \geq -K(c_{j+1}, \ldots, c_{j+k}; \underline{m}_{j+1}, \ldots, \underline{m}_{j+k}; i_{j+1}, \ldots, i_{j+k}; f_0) + \sum_{r=j+1}^{j+k} c_r \alpha_r,$$

with equality iff f *is the* GvMF_{j+k} *density given in (a).*

4.3 Other Properties

In this section we first provide some other properties of GvMF distributions, which concern the invariance, the clustering around the mode and the antipodal symmetry, and we show a few of the various possible shapes taken by GvMF densities. Then we show various relationships between GvMF and other directional or linear distributions.

The first property of GvMF distributions is their invariance under orthogonal transformations. If $\underline{X} \sim \mathrm{GvMF}_k(\underline{m}_1, \ldots, \underline{m}_k; c_1, \ldots, c_k; i_1, \ldots, i_k)$ on S^{p-1} and A is a $p \times p$ orthogonal matrix, then $\underline{Y} = A\underline{X} \sim \mathrm{GvMF}_k(A\underline{m}_1, \ldots, A\underline{m}_k; c_1, \ldots, c_k; i_1, \ldots, i_k)$, meaning that the class of GvMF_k distributions with fixed $c_1, \ldots, c_k$ and $i_1, \ldots, i_k$ is invariant under the group of orthogonal transformations on R^p. As practical consequence, the choice of coordinate system is irrelevant.

Regarding the shapes taken by the GvMF densities we can provide the following general statements. If i_r is even, then the clustering around both directions $\underline{m}_r$ and $-\underline{m}_r$ increases as c_r increases, $r = 1, \ldots, k$. If $i_1, \ldots, i_k$ are even, then $(dP)/(dU)(\underline{x}) = (dP)/(dU)(-\underline{x})\ \forall \underline{x} \in S^{p-1}$, meaning that we have invariance under $\underline{x} \mapsto -\underline{x}$, $\underline{x} \in S^{p-1}$, i.e., antipodal symmetry, and we have thus a distribution for axial data. In general, the larger the values of $c_1, \ldots, c_k$, the larger the

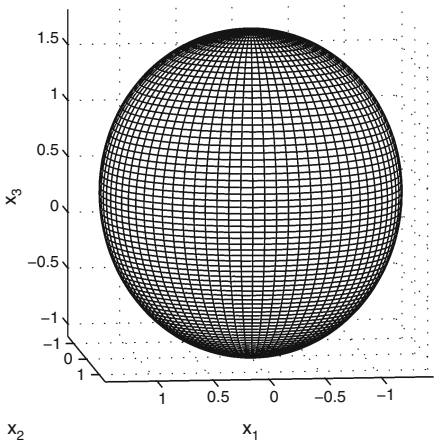


Fig. 4.1 $\mathrm{GvMF}_1((0,0,1);1;1)$

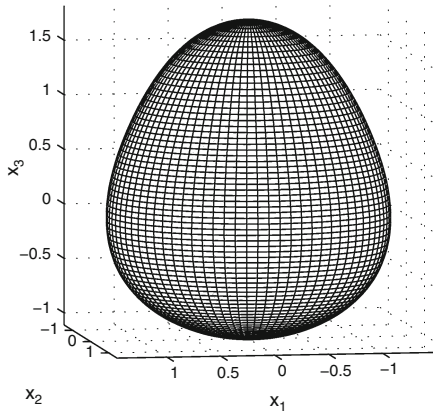


Fig. 4.2 $\mathrm{GvMF}_1((0,0,1);1;3)$

clustering around $\underline{m}_1, \ldots, \underline{m}_k$. If $\underline{m}_1 = \ldots = \underline{m}_k$ and $c_1, \ldots, c_k > 0$, then the maximum of the density is at $\underline{m}_1$ and the minimum at $-\underline{m}_1$. If $\underline{m}_1 = \ldots = \underline{m}_k$, then we have the rotationally symmetry around $\underline{m}_1$. Clearly, the GvMF_k distribution with $c_1 = \ldots = c_k = 0$ is the isotropic distribution. Any GvMF_k distribution with $i_1 = \ldots = i_k = 1$ is simply the $\mathrm{vMF}(||\underline{\nu}||^{-1}\underline{\nu}; ||\underline{\nu}||)$ distribution, whenever $\underline{\nu} = \sum_{r=1}^{k} c_r \underline{m}_r \neq 0$.

Figures 4.1–4.12 show some particular GvMF densities on S^2. The densities are plotted around the unit sphere and up to unspecified proportionality constants. GvMF_1 densities with $i_1 = 1, 3$ and 7 are respectively plotted in Figs. 4.1, 4.2 and 4.3, all other parameters being held fixed. These figures show hence the effect of the increase of i_1 in the GvMF_1 density. Figures 4.2 and 4.4 have GvMF_1 densities with $c_1 = 1$ and 2 respectively, all other parameters being equal. Figures 4.4 and 4.5 are GvMF_1 densities with $i_1 = 3$ and 4 respectively, all other parameters being equal. Both Figs. 4.5 and 4.11 show antipodal symmetries, yielding distributions for axial data. Figure 4.5 has $i_1 = 4$ even and has hence a mode and a similar antimode.

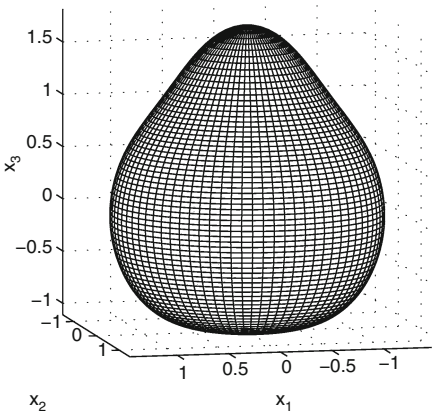


Fig. 4.3 $\mathrm{GvMF}_1((0,0,1);1;7)$

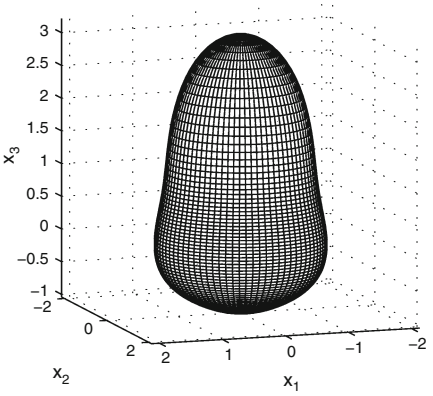


Fig. 4.4 $\mathrm{GvMF}_1((0,0,1);2;3)$

Figure 4.11 is a GvMF_3 density with $i_1 = i_2 = i_3 = 4$ and hence with three modes and three similar antimodes. Due to the choice of the other parameters, all six modes are in fact similar. Figures 4.6 and 4.7 show two GvMF_2 densities and Figs. 4.8–4.11 show four GvMF_3 densities. Finally, Fig. 4.12 shows a girdle shaped GvMF_4 density.

Another existing generalization of the GvMF distribution is the Fisher–Bingham distribution, whose density at $\underline{x} \in S^{p-1}$ is proportional to

$$\exp\{c\underline{m}^{\mathrm{T}}\underline{x} + \underline{x}^{\mathrm{T}}M\underline{x}\},$$

where $c \geq 0$, $\underline{m} \in S^{p-1}$ and M is a $p \times p$ symmetric matrix standardized to have null trace (see, e.g., [10, p. 174]). The restriction to matrices with null trace affects the normalizing constant only. If we restrict M to have rank one and relax the null trace condition, then both symmetry and unitary rank conditions are equivalent to the existence of a vector $\underline{a} \in \mathrm{R}^p$ such that $M = \underline{a}\underline{a}^{\mathrm{T}}$. By setting $\underline{a} = \sqrt{c_2}\underline{m}_2$, where $c_2 \geq 0$ and $||\underline{m}_2|| = 1$, we see that the GvMF_2 distribution with $i_1 = 1$ and $i_2 = 2$ is the special case of the Fisher–Bingham distribution with $\mathrm{rank}\{M\} = 1$.

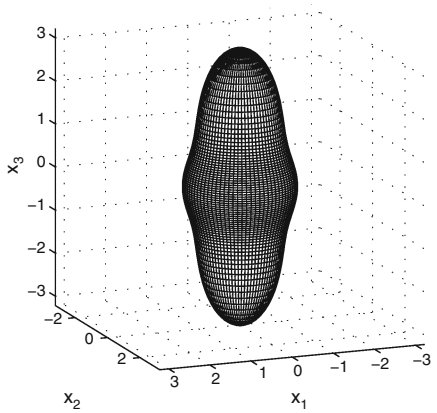


Fig. 4.5 $\mathrm{GvMF}_1((0,\ 0,\ 1);\ 2;\ 4)$

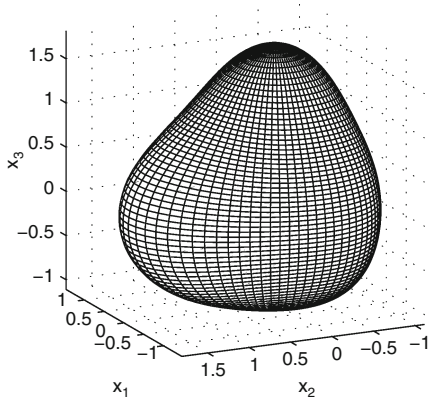


Fig. 4.6 $\mathrm{GvMF}_2((0, 1, 0),\ (0, 0, 1);\ 1, 1;\ 3, 5)$

This latter GvMF_2 distribution is in fact known as Fisher–Watson distribution (see [10, p. 177]).

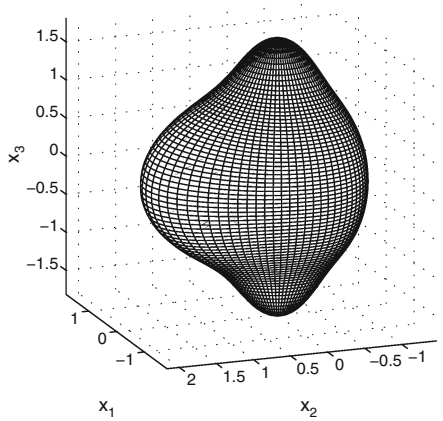


Fig. 4.7 $\mathrm{GvMF}_2((0, 1, 0),\ (0, 0, 1);\ 1, 1;\ 3, 10)$

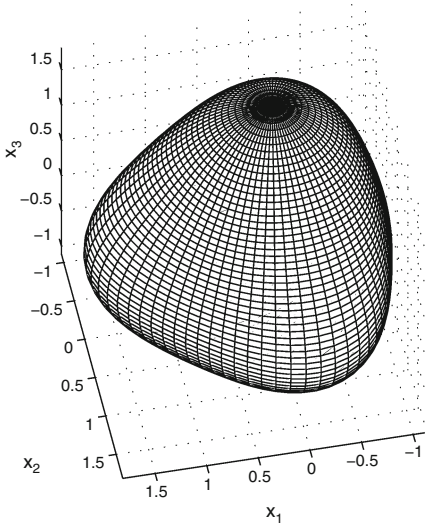


Fig. 4.8 $\mathrm{GvMF}_3((1,0,0),(0,1,0),(0,0,1);1,1,1;3,3,3)$

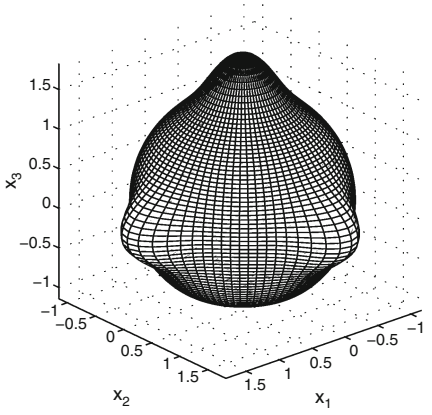


Fig. 4.9 $\mathrm{GvMF}_3((1,0,0),(0,1,0),(0,0,1);1,1,1;15,15,15)$

For $p = 2$, a broad class of circular distributions which plays an important role w.r.t. the information theoretic quantities defined in Sect. 4.2 is the generalized von Mises of order k (GvM_k). In terms of the angular coordinate and w.r.t. isotropic or the Lebesgue measure on S^1, the GvM_k density is given by

$$f(\theta \mid \mu_1,\ldots,\mu_k;\kappa_1,\ldots,\kappa_k) = \frac{1}{2\pi G_0^{(k)}(\delta_1,\ldots,\delta_{k-1};\kappa_1,\ldots,\kappa_k)} \exp\left\{\sum_{r=1}^{k}\kappa_r \cos r(\theta-\mu_r)\right\}, \quad (4.12)$$

for $\theta \in [0,2\pi)$, $\kappa_r \geq 0$, $\mu_r \in [0,2\pi/r)$, for $r = 1,\ldots,k$, and where

Fig. 4.10 $\mathrm{GvMF}_3((1,0,0),(0,1,0),(0,0,1);1,1,1;15,15,15)$

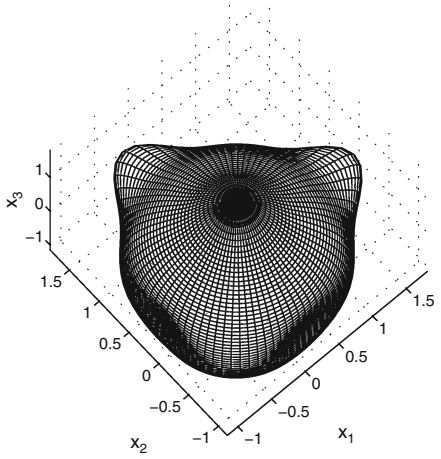


Fig. 4.11 $\mathrm{GvMF}_3((0,0,1),(0,1,0),(0,0,1);2,2,2;4,4,4)$

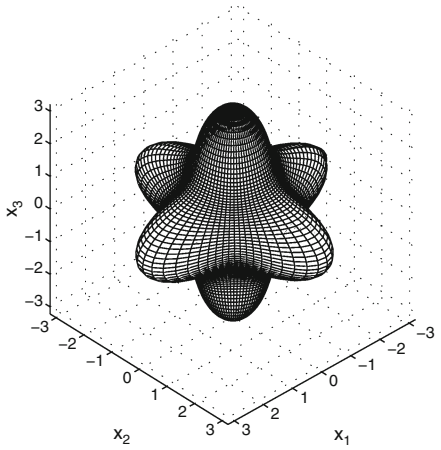
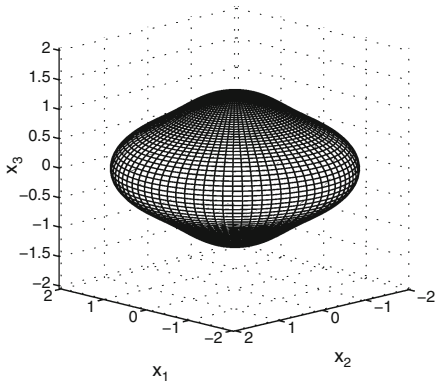


Fig. 4.12 $\mathrm{GvMF}_4((0,1,0),(1,1,0)/\sqrt{2},(1,0,0),(1,-1,0)/\sqrt{2};1,1,1,1;6,6,6,6)$



$$G_0^{(k)}(\delta_1,\ldots,\delta_{k-1};\kappa_1,\ldots,\kappa_k) = \frac{1}{2\pi}\int_0^{2\pi} \exp\{\kappa_1\cos\theta + \kappa_2\cos 2(\theta+\delta_1) + \ldots + \kappa_k\cos k(\theta+\delta_{k-1})\}d\theta,$$

with $\delta_1 = (\mu_1-\mu_2)\mathrm{mod}\,\pi, \delta_2 = (\mu_1-\mu_3)\mathrm{mod}(2\pi/3), \ldots, \delta_{k-1} = (\mu_1-\mu_k)\mathrm{mod}\,(2\pi/k)$. We denote by $\mathrm{GvM}_k(\mu_1,\ldots,\mu_k;\kappa_1,\ldots,\kappa_k)$ a representative angle with the density (4.12). This class of distributions was introduced by [9] and [5] as a generalization of the circular normal or von Mises (vM) distribution, whose density is given by

$$f(\theta \mid \mu;\kappa) = \frac{1}{2\pi I_0(\kappa)}\exp\{\kappa\cos(\theta-\mu)\},$$

for $\theta,\mu \in [0,2\pi)$, $\kappa > 0$, where $I_n(z) = (2\pi)^{-1}\int_0^{2\pi}\cos n\theta\,\exp\{z\cos\theta\}d\theta$, $z \in \mathrm{C}$, is the modified Bessel function of the first kind and integer order n (see, e.g., [1, p. 376]). Compared to the vM, which is circular symmetric and unimodal, the GvM_2 distribution allows already for higher flexibility in terms of asymmetry and bimodality.

When $p = 2$, many GvMF_k densities with are indeed GvM_k densities. The following result we illustrates this fact for the case $k = 3$ and $i_r = r$, $r = 1,2,3$.

Result 4.1. Consider $\underline{X} \sim \mathrm{GvMF}_3(\underline{m}_1,\underline{m}_2,\underline{m}_3;c_1,c_2,c_3;1,2,3)$ on S^1 and define $\theta = \arg\{\underline{X}\}$. Then $\theta \sim \mathrm{GvM}_3(\mu_1,\mu_2,\mu_3;\kappa_1,\kappa_2,\kappa_3)$, where for $\nu_r = \arg\{\underline{m}_r\}$, $r = 1,2,3$, and $\arctan : \mathrm{R} \to (-\pi/2,\pi/2)$, we have

$$\mu_1 = \left(\arctan\frac{4c_1\sin\nu_1 + 3c_3\sin\nu_3}{4c_1\cos\nu_1 + 3c_3\cos\nu_3} + \pi I\{4c_1\cos\nu_1 + 3c_3\cos\nu_3 < 0\}\right)\mathrm{mod}\,2\pi, \tag{4.13}$$

$$\kappa_1 = \{c_1^2 + (9/16)c_3^2 + (3/2)c_1c_3\cos(\nu_1-\nu_3)\}^{\frac{1}{2}}, \tag{4.14}$$

$\mu_2 = \nu_2, \kappa_2 = c_2/2, \mu_3 = \nu_3$ and $\kappa_3 = c_3/4$.

Proof. Let us define $\theta = \arg\{\underline{x}\}$, where $\underline{x} \in S^1$, $\nu_r = \arg\{\underline{m}_r\}$ and $\theta_r = \theta - \nu_r$, for $r = 1,2,3$. From $\underline{m}_1^{\mathrm{T}}\underline{x} = \cos\theta_1$, $(\underline{m}_2^{\mathrm{T}}\underline{x})^2 = \cos^2\theta_2 = (1/2)\cos 2\theta_2 + 1/2$ and $(\underline{m}_3^{\mathrm{T}}\underline{x})^3 = \cos^3\theta_3 = (1/4)\cos 3\theta_3 + (3/4)\cos\theta_3$ and from the fact that the transformation from Cartesian to polar coordinates has Jacobian one, it follows from (4.1) that the density of $\arg\{\underline{X}\}$ at θ is proportional to

$$\exp\left\{c_1\cos(\theta-\nu_1) + \frac{3}{4}c_3\cos(\theta-\nu_3) + \frac{1}{2}c_2\cos 2(\theta-\nu_2) + \frac{1}{4}c_3\cos 3(\theta-\nu_3)\right\}. \tag{4.15}$$

The first two summands of the above exponent can be re-written as

$$\cos\theta\left(c_1\cos\nu_1 + \frac{3}{4}c_3\cos\nu_3\right) + \sin\theta\left(c_1\sin\nu_1 + \frac{3}{4}c_3\sin\nu_3\right)$$

$$= \cos\theta\,(\kappa_1\cos\mu_1) + \sin\theta\,(\kappa_1\sin\mu_1) = \kappa_1\cos(\theta-\mu_1),$$

with μ_1 and κ_1 given by (4.13) and (4.14). From this follows the GvM_3 density at θ

$$\frac{1}{2\pi G_0^{(3)}(\delta_1,\delta_2;\kappa_1,\kappa_2,\kappa_3)}\exp\{\kappa_1\cos(\theta-\mu_1)+\kappa_2\cos 2(\theta-\mu_2)+\kappa_3\cos 3(\theta-\mu_3)\},$$

where μ_1 is given by (4.13), κ_1 is given by (4.14), $\mu_2 = \nu_2, \kappa_2 = c_2/2, \mu_3 = \nu_3, \kappa_3 = c_3/4, \delta_1 = (\mu_1-\mu_2)\mathrm{mod}\,\pi$ and $\delta_2 = (\mu_1-\mu_3)\mathrm{mod}(2\pi/3)$. If $\nu_1 = \nu_3\mathrm{mod}\,\pi$ then the cosine in (4.14) is equal to -1 and $\kappa_1 = |c_1 - 3/4\,c_3|$. Hence the square root in (4.14) is always real. If $\nu_1 = \nu_3\mathrm{mod}\,2\pi$, then $\kappa_1 = c_1 + 3/4\,c_3$, which is in fact clear from (4.15). □

In this context note that a series expansion for the normalizing constant $G_0^{(3)}$ is provided by (24) in [4]. The precise relationships between the GvMF_k and the GvM_k distributions are straightforward to obtain for the cases $i_1 = 1, k = 1$ and $i_1 = 1, i_2 = 2, k = 2$. The assertion that there is an analogue relationship between GvMF_k and GvM_k densities for $1 \le i_k \le k$ and $k > 3$ follows from the fact that $\cos^k$ can always be expanded into a sum of multiple angle cosines with multiples lower than or equal to k. For example, $\cos^5\alpha = (\cos 5\alpha + 5\cos 3\alpha + 10\cos\alpha)/16$. Hence, it must always be possible to re-express a GvMF_k density (4.1) with $1 \le i_k \le k$ in terms of the GvM_k density (4.12). However, general formulae for the GvM_k parameters in terms if the GvMF_k parameters seem difficult to establish for $k > 3$. So GvMF and GvM distributions are closely related and in fact, Theorem 4.1 and Corollaries 4.1 and 4.2 are the analogues of Theorem 2.1 and Corollaries 2.1 and 2.2 by [4], respectively, of the GvM distribution.

The next results provide characterizations of GvMF distributions in terms of well known multivariate distributions, Result 4.2 is a characterization in terms of the multivariate normal distribution. Let us remind that a conditional offset distribution is the conditional distribution of a random vector given that it has norm equal to one.

Result 4.2. Assume $\underline{X} \sim \mathcal{N}(\underline{\mu}, \Sigma)$ in R^p, with Σ positive definite. If $\underline{\mu} \neq 0$, then $\underline{X}$ has a GvMF_{p+1} conditional offset distribution on S^{p-1}. More precisely,

$$\underline{X} \mid ||\underline{X}|| = 1 \sim \mathrm{GvMF}_{p+1}(\underline{m}_1,\ldots,\underline{m}_{p+1}; c_1,\ldots,c_{p+1}; 1,2,\ldots,2),$$

where, given $V = (\underline{v}_{\circ 1},\ldots,\underline{v}_{\circ p})$ nonsingular such that $\Sigma^{-1} = VV^{\mathrm{T}}$, we have $\underline{m}_1 = ||\Sigma^{-1}\underline{\mu}||^{-1}\Sigma^{-1}\underline{\mu}$, $c_1 = ||\Sigma^{-1}\underline{\mu}||$, $\underline{m}_{j+1} = ||\underline{v}_{\circ j}||^{-1}\underline{v}_{\circ j}$, $c_{j+1} = -||\underline{v}_{\circ j}||^2/2$, for $j = 1,\ldots,p$. If $\underline{\mu} = 0$, then

$$\underline{X} \mid ||\underline{X}|| = 1 \sim \mathrm{GvMF}_p(\underline{m}_2,\ldots,\underline{m}_{p+1}; c_2,\ldots,c_{p+1}; 2,\ldots,2).$$

Proof. The conditional random vector $\underline{X} \mid ||\underline{X}|| = 1$ has density at $\underline{x} \in S^{p-1}$ proportional to

$$\exp\left\{\underline{\mu}^{\mathrm{T}}\Sigma^{-1}\underline{x} - \frac{1}{2}\underline{x}^{\mathrm{T}}\Sigma^{-1}\underline{x}\right\}. \tag{4.16}$$

Clearly, $\underline{\mu}^{\mathrm{T}}\Sigma^{-1}\underline{x} = c_1\underline{m}_1^{\mathrm{T}}\underline{x}$. If $\underline{\mu} \neq 0$, then we can define $\underline{m}_1$ and c_1 as in Result 4.2. If $\underline{\mu} = 0$, then the term $c_1\underline{m}_1^{\mathrm{T}}\underline{x}$ vanishes from the GvMF_{p+1} density. The three following statements are equivalent: Σ is positive definite, Σ^{-1} is positive definite and, third, there exists V nonsingular such that $\Sigma^{-1} = VV^{\mathrm{T}}$. As $\underline{v}_{\circ j}$ denotes the jth column of V, $j = 1, \ldots, p$, we have $VV^{\mathrm{T}} = \sum_{j=1}^{p} \underline{v}_{\circ j}\underline{v}_{\circ j}^{\mathrm{T}}$. With this,

$$
\begin{aligned}
-\frac{1}{2}\underline{x}^{\mathrm{T}}\Sigma^{-1}\underline{x} &= -\frac{1}{2}\underline{x}^{\mathrm{T}}\sum_{j=1}^{p}\underline{v}_{\circ j}\underline{v}_{\circ j}^{\mathrm{T}}\underline{x} = \sum_{j=1}^{p} -\frac{1}{2}||\underline{v}_{\circ j}||^2\left(\frac{\underline{v}_{\circ j}^{\mathrm{T}}\underline{x}}{||\underline{v}_{\circ j}||}\right)^2 \\
&= \sum_{j=1}^{p} c_{j+1}(\underline{m}_{j+1}^{\mathrm{T}}\underline{x})^2,
\end{aligned}
$$

which terminates the proof. Note that because V is nonsingular, $\underline{v}_{\circ 1}, \ldots, \underline{v}_{\circ p}$ are linearly independent and the same holds for $\underline{m}_2, \ldots, \underline{m}_{p+1}$. □

Result 4.3 below extends the characterization of the GvMF distribution given by Result 4.2 to the situation where the covariance matrix is singular.

Result 4.3. Let Σ be a $p \times p$ semi-positive definite matrix of rank $r < p$ and denote by $\lambda_1, \ldots, \lambda_r > 0$ its non-zero eigenvalues. Let $\underline{X} \in \mathrm{R}^p$ have the singular normal density at $\underline{x} \in \mathrm{R}^p$ given by

$$
(2\pi)^{-\frac{r}{2}}(\lambda_1 \cdot \ldots \cdot \lambda_r)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu})^{\mathrm{T}}\Sigma^{-}(\underline{x}-\underline{\mu})\right\}, \tag{4.17}
$$

where Σ^{-} is a generalized inverse of Σ, i.e., a matrix satisfying $\Sigma\Sigma^{-}\Sigma = \Sigma$. Then, for $\underline{\mu} \neq 0$, $\underline{X}$ has the $\mathrm{GvMF}_{r+1}(\underline{m}_1, \ldots, \underline{m}_{r+1}; c_1, \ldots, c_{r+1}; 1, 2, \ldots, 2)$ conditional offset distribution on S^{p-1}, where the parameters $\underline{m}_j, c_j$, for $j = 1, \ldots, r+1$, have the same form as under Result 4.2, with Σ^{-} replacing Σ^{-1}. If $\underline{\mu} = 0$, then $\underline{X}$ has a $\mathrm{GvMF}_r(\underline{m}_2, \ldots, \underline{m}_{r+1}; c_2, \ldots, c_{r+1}; 2, \ldots, 2)$ conditional offset distribution on S^{p-1}.

Proof. For $\Lambda_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$ and $\Gamma_1 = (\underline{\gamma}_{\circ 1}, \ldots, \underline{\gamma}_{\circ r})$, where $\underline{\gamma}_{\circ j}$ is the normed eigenvector of Σ associated to λ_j, $j = 1, \ldots, r$, we have $\Sigma^{-} = \Gamma_1\Lambda_1^{-1}\Gamma_1^{\mathrm{T}}$ (see, e.g., [11, p. 477]). Define $V = (\underline{v}_{\circ 1}, \ldots, \underline{v}_{\circ r}) = \Gamma_1\Lambda_1^{-1/2}$. Then $\Sigma^{-} = VV^{\mathrm{T}} = \sum_{j=1}^{r}\underline{v}_{\circ j}\underline{v}_{\circ j}^{\mathrm{T}}$. As the conditional offset density at $\underline{x} \in S^{p-1}$ is proportional to

$$
\exp\left\{\underline{\mu}^{\mathrm{T}}\Sigma^{-}\underline{x} - \frac{1}{2}\underline{x}^{\mathrm{T}}\Sigma^{-}\underline{x}\right\},
$$

it follows that $\underline{\mu}^{\mathrm{T}}\Sigma^{-}\underline{x} = c_1\underline{m}_1^{\mathrm{T}}\underline{x}$ and

$$-\frac{1}{2}\underline{x}^{\mathrm{T}}\Sigma^{-}\underline{x} = -\frac{1}{2}\underline{x}^{\mathrm{T}}\sum_{j=1}^{r}\underline{v}_{\circ j}\underline{v}_{\circ j}^{\mathrm{T}}\underline{x} = \sum_{j=1}^{r}-\frac{1}{2}||\underline{v}_{\circ j}||^2\left(\frac{\underline{v}_{\circ j}^{\mathrm{T}}\underline{x}}{||\underline{v}_{\circ j}||}\right)^2$$
$$= \sum_{j=1}^{r} c_{j+1}(\underline{m}_{j+1}^{\mathrm{T}}\underline{x})^2.$$

□

Note that the density (4.17) of $\underline{X}$ is defined on the translated sub-vector space $\{\underline{x} \in \mathrm{R}^p | N^{\mathrm{T}}(\underline{x} - \underline{\mu}) = 0\}$, where N is a $p \times (p - r)$ matrix satisfying $N^{\mathrm{T}}\Sigma = 0$ and $N^{\mathrm{T}}N = I_{p-r}$. Given $\Sigma = \Gamma\Lambda\Gamma^{\mathrm{T}}$, with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)$ and $\Gamma = (\Gamma_1, \Gamma_2)$ with orthogonal and normed columns, $\underline{X}$ can be expressed as $\underline{X} = \Gamma_1\underline{Y} + \underline{\mu}$, where $\underline{Y} \sim \mathcal{N}(0, \Lambda_1)$. Hence N can be taken equal to Γ_2. For further details refer e.g. to [11, p. 42].

Gatto and Jammalamadaka [5, Characterization 2, p. 348], shows that for the case $p = 2$, the conditional offset distribution of a bivariate normal vector is a GvM_2 distribution. As noted by Result 4.1 and by the paragraph following it, any GvM_2 distribution can be re-expressed as a GvMF_2 distribution, and hence Result 4.2 generalizes this characterization of the GvM_2 distribution to dimensions $p > 2$. Result 4.2 with $\underline{\mu} = 0$ yields the GvMF_p conditional offset distribution having p quadratic summands only. In agreement with this, when $p = 2$, we can see from [5, Characterization 2], that $\underline{\mu} = 0$ leads to the bimodal von Mises conditional offset distribution, i.e., to the GvM_2 density given in (4.12) with $\kappa_1 = 0$.

We can also deduce directly from (4.16) that if $\Sigma = \sigma^2 I$, for $\sigma > 0$, and $\underline{\mu} \neq 0$, then $\underline{X} \,|\, ||\underline{X}|| = 1 \sim \mathrm{vMF}(||\underline{\mu}||^{-1}\underline{\mu}; \sigma^{-2}||\underline{\mu}||)$, which is well known.

The next result shows that the GvMF distribution arises also as conditional offset distribution from the exponential spherical class suggested by [3] as a generalization of the multivariate normal distribution. This class consists of multivariate distributions with densities at $\underline{x} \in \mathrm{R}^p$ proportional to

$$\exp\left\{-\frac{1}{r}[(\underline{x} - \underline{\lambda})^{\mathrm{T}}W(\underline{x} - \underline{\lambda})]^{\frac{r}{2}}\right\}, \tag{4.18}$$

where $\underline{\lambda} \in \mathrm{R}^p$, W is a $p \times p$ positive definite matrix and $r > 1$.

Result 4.4. Assume $\underline{X} \in \mathrm{R}^p$ has the exponential spherical density proportional to (4.18) with $W = wI$, for $w > 0$, and $k = r/2 \in \{1, 2, \ldots\}$. If $\underline{\lambda} \neq 0$, then $\underline{X}$ has a GvMF_k conditional offset distribution on S^{p-1}. More precisely,

$$\underline{X} \,|\, ||\underline{X}|| = 1 \sim \mathrm{GvMF}_k(\underline{m}, \ldots, \underline{m}; c_1, \ldots, c_k; 1, \ldots, k),$$

where $\underline{m} = ||\underline{\lambda}||^{-1}\underline{\lambda}$ and

$$c_r = (-2)^{r-1}\frac{w^k}{k}\binom{k}{j}(1 + ||\underline{\lambda}||^2)^{k-r}||\underline{\lambda}||^r, \tag{4.19}$$

for $r = 1, \ldots, k$. If $\underline{\lambda} = 0$, then the conditional offset distribution of $\underline{X}$ is the isotropic one.

Proof. The conditional random vector $\underline{X} \,|\, ||\underline{X}|| = 1$ has density at $\underline{x} \in S^{p-1}$ proportional to

$$\exp\left\{-\frac{w^k}{2k}[(\underline{x}-\underline{\lambda})^{\mathrm{T}}(\underline{x}-\underline{\lambda})]^k\right\}.$$

The above exponent can be expanded as

$$-\frac{w^k}{2k}\sum_{r=0}^{k}\binom{k}{r}(-2\underline{\lambda}^{\mathrm{T}}\underline{x})^r(1+\underline{\lambda}^{\mathrm{T}}\underline{\lambda})^{k-r},$$

which is, up to an additive constant and for $\underline{\lambda} \neq 0$, equal to $\sum_{r=1}^{k} c_r(\underline{m}^{\mathrm{T}}\underline{x})^r$, with $\underline{m}$ and c_r as given above, for $r = 1, \ldots, k$. □

From Result 4.1, the GvMF_3 distribution can be reparametrized in the GvM_3 form when $p = 2$ and thus Result 4.4 extends in some way the characterization of the GvM_2 distribution in terms of the bivariate normal distribution given in [5, Characterization 2]. The case $k = 3$ is given precisely in Result 4.5 below.

Result 4.5. Assume $\underline{X} \in \mathrm{R}^2$ has a density proportional to (4.18) with $W = wI$, for $w > 0$ and $k = r/2 = 3$. Then

$$\arg\{\underline{X}\} \,|\, ||\underline{X}|| = 1 \sim \mathrm{GvM}_3(\mu_1, \mu_2, \mu_3; \kappa_1, \kappa_2, \kappa_3),$$

where for c_1, c_2 and c_3 given by (4.19), μ_1 is given by (4.13) and κ_1 by (4.14) with $\nu_1 = \nu_3 = \arg\{\underline{\lambda}\}$, $\mu_2 = \mu_3 = \arg\{\underline{\lambda}\}$, $\kappa_2 = c_2/2$ and $\kappa_3 = c_3/4$.

Acknowledgements The author is grateful to Professor Sreenivasa Rao Jammalamadaka, for many constructive and stimulating discussions in the field of directional statistics, to the Editors and to the Referees.

References

1. Abramowitz M, Stegun IE (1972) Handbook of mathematical functions with formulas, graphs, and mathematical tables, 9th printing. Dover, New York. Originally published by the National Bureau of Standards, USA, 10th printing
2. Beran R (1979) Exponential models for directional data. Ann Stat 7:1162–1178
3. de Simoni S (1968) Su una estensione dello schema delle curve normali di ordine r alle variabili doppie. Statistica 37:447–474
4. Gatto R (2009) Information theoretic results for circular distributions. Statistics (Ber) 43: 409–421
5. Gatto R, Jammalamadaka SR (2007) The generalized von Mises distribution. Stat Methodol 4:341–353

6. Jammalamadaka SR, SenGupta A (2001) Topics in circular statistics. World Scientific, Singapore
7. Kullback S (1954) Certain inequalities in information theory and the Cramer–Rao inequality. Ann Math Statist 25:745–751
8. Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Statist 22:79–86
9. Maksimov VM (1967) Necessary and sufficient statistics for the family of shifts of probability distributions on continuous bicompact groups. Theoriya Veroyatnostei i ee Primeneniya, 12:307–321 (in Russian); Theory Probab Appl 12:267–280 (English translation)
10. Mardia KV, Jupp PE (2000) Directional statistics. Wiley, Chichester
11. Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. Academic, London
12. Shannon CE (1948) A mathematical theory of communication. Bell Sys Tech J 27:379–423, 623–656

Chapter 5
A New Nonparametric Test of Symmetry

Kaushik Ghosh

Abstract We present a new nonparametric test for symmetry about an unknown location and investigate its large sample properties. Asymptotic normality of the test statistic is established and an estimator of the asymptotic variance is also presented. Results of a simulation study and data analysis are also presented.

5.1 Introduction

In many practical situations, it is important to determine whether the underlying population has a symmetric distribution. For example, if one is interested in estimating the measure of location, having a skewed distribution would give rise to consideration of more than one such measure. In the case of a paired sample problem with "treatment" and "control", the hypothesis of "no treatment effect" implies that the paired difference $Z = X - Y$ is symmetric about 0. Many robust statistical methods (see [35]) depend on the assumption of symmetry. In case symmetry is not valid, one would need to determine a symmetrizing transformation before applying the statistical procedures. Koziol [40] mentions using a test for symmetry as a screening procedure before applying the modulus family of transformations introduced by [36] for bringing data to be closer to normality.

The problem of testing univariate symmetry has received attention in the literature for quite some time. Past work in this area can be broadly classified into two groups – one where the center of symmetry is known (without loss of generality, this is then taken to be 0) and the other case is where the center of symmetry is completely unknown. A common class of nonparametric tests in the first category is the weighted sign test, which has been studied in detail by many authors, such as [29, 30, 37, 59]. Special cases of the former include

K. Ghosh
Department of Mathematical Sciences, University of Nevada at Las Vegas, Box 454020, 4505 Maryland Parkway, Las Vegas, NV 89154-4020, USA
e-mail: kaushik.ghosh@unlv.edu

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_5,

the popular sign test and the signed-rank test. Other examples include works by [14, 20, 47, 53, 56]. Aki [3] proposed a new test of symmetry which generalized the results of [14] and [53]. Ahmad [1] proposed asymptotically normal estimates of symmetry statistic by [53]. Reynolds [51] and Rosenstein [52] developed a sequential signed-rank test for symmetry about a known value. Further work have been done by [6, 8, 9, 17, 21, 31–33, 39, 41, 42]. Some of the recent work in this area have been by [2, 15, 45, 46, 58].

The case of unknown center of symmetry is a more realistic and practical problem. One of the earliest articles in this area is by [34] where the authors use the difference of the mean and median as a measure of skewness. One group of tests is obtained by applying the tests with known center, where the center is estimated from the data. Gastwirth [26] proposed the modified sign test, whereby the unknown point of symmetry is estimated by the sample mean and a sign test is developed for symmetry around this point. Other articles in this category include those by [10, 21, 28]. Boos [13] propose a test for symmetry based on the Hodges–Lehman estimator of location. Davis and Quade [19] and Randles [50] propose asymptotically distribution free tests of symmetry. Antille and Kersting [5], Antille et al. [6] and Finch [25] propose tests based on symmetric differences of spacings. Csörgő and Heathcote [18] and Koutrouvelis [38] use the empirical characteristic function to develop tests for symmetry about an unknown location, that do not require estimation of the unknown location. Koziol [40] propose using rank-based tests of symmetry with known center where the center is estimated using asymptotically efficient methods. Eubank et al. [23] also propose a class of rank-based tests of symmetry. Schuster and Barker [54] proposed a symmetric bootstrap procedure for testing symmetry about unknown center, which was later examined in further detail by [7] and extended to the dependent case by [48]. Cabilio and Masaro [16] propose a test of symmetry about an unknown location based on deviation of the sample mean from the sample median, similar to the lines of [34]. Some of the recent work in this area are [22, 43, 44].

In this article, we develop a new test of symmetry for univariate distributions when the center of symmetry is unknown. Our test is based on a new characterization of symmetry. This article proceeds as follows. In Sect. 5.2, we propose the new estimator and investigate its properties. In Sect. 5.3, we provide results of simulation studies in support of the theoretical properties. In Sect. 5.4, we provide a real-life example of use of the proposed statistic. Finally, we end with concluding remarks in Sect. 5.5. The proofs of results presented here are presented in the Appendix.

5.2 Theory

Suppose $X_1, \ldots, X_n$ is a set of observations (not necessarily i.i.d.) from an underlying continuous distribution with cdf F and corresponding pdf f. We want to test for symmetry of f about an unknown point μ (say). For any $p \in (0,\ 1)$, consider the function

$$g(p) = F\left(\frac{F^{-1}(p) + F^{-1}(1-p)}{2}\right).$$

Note that $g(p) < \frac{1}{2}$ if f is left skewed, $g(p) > \frac{1}{2}$ if f is right skewed and $g(p) = \frac{1}{2}$ is f is symmetric. Also note that $g(p) = g(1-p)$. Consequently, we propose the following quantity to characterize the symmetry of f:

$$\tau \equiv \tau(F) = \int_0^{1/2} g(p)dp. \tag{5.1}$$

It follows from prior remarks that $\tau = \frac{1}{4}$ for symmetric, $\tau > \frac{1}{4}$ for right skewed and $\tau < \frac{1}{4}$ for left skewed distributions.

Suppose $\hat{F}$ is an empirical-process based estimate of F, such that

$$\sqrt{n}(\hat{F} - F) \xrightarrow{d} W_F, \tag{5.2}$$

where W_F is a Gaussian process with covariance kernel $K_F(\cdot,\cdot)$. We propose estimating τ using the quantity

$$\hat{\tau} = \int_0^{1/2} \hat{F}\left(\frac{\hat{F}^{-1}(p) + \hat{F}^{-1}(1-p)}{2}\right) dp. \tag{5.3}$$

Our goal is now to estimate the asymptotic distribution of $\hat{\tau}$. Consider the map $\phi : D[0,\ 1] \to D[0,\ 1]$ given by

$$\phi(F)(p) = F\left(\frac{F^{-1}(p) + F^{-1}(1-p)}{2}\right). \tag{5.4}$$

We first present an important result.

Theorem 5.1. *Suppose F is a distribution with pdf f and $\hat{F}$ be an empirical-process based estimate of F satisfying (5.2), where W_F is a Gaussian process with covariance kernel $K_F(\cdot,\cdot)$. Suppose we define the functional $G = \phi(F)$ and $\hat{G}$ be its empirical counterpart, where ϕ is given by (5.4). We then have*

$$\sqrt{n}(\hat{G} - G) \xrightarrow{d} W_G,$$

where W_G is a Gaussian process with covariance kernel $K_G(\cdot,\cdot)$ given by

$$K_G(s,t) = K_F\left(\frac{F^{-1}(s) + F^{-1}(1-s)}{2}, \frac{F^{-1}(t) + F^{-1}(1-t)}{2}\right)$$
$$-\frac{1}{2} f\left(\frac{F^{-1}(s) + F^{-1}(1-s)}{2}\right)\left\{\frac{1}{fF^{-1}(s)} K_F\left(F^{-1}(s), \frac{F^{-1}(t)+F^{-1}(1-t)}{2}\right)\right.$$

$$
\begin{aligned}
&+\frac{1}{fF^{-1}(1-s)}K_F\left(F^{-1}(1-s),\frac{F^{-1}(t)+F^{-1}(1-t)}{2}\right)\Bigg\}\\
&-\frac{1}{2}f\left(\frac{F^{-1}(t)+F^{-1}(1-t)}{2}\right)\Bigg\{\frac{1}{fF^{-1}(t)}K_F\left(F^{-1}(t),\frac{F^{-1}(s)+F^{-1}(1-s)}{2}\right)\\
&+\frac{1}{fF^{-1}(1-t)}K_F\left(F^{-1}(1-t),\frac{F^{-1}(s)+F^{-1}(1-s)}{2}\right)\Bigg\}\\
&+\frac{1}{4}f\left(\frac{F^{-1}(s)+F^{-1}(1-s)}{2}\right)f\left(\frac{F^{-1}(t)+F^{-1}(1-t)}{2}\right)\\
&\times\Bigg\{\frac{1}{fF^{-1}(s)fF^{-1}(t)}K_F\left(F^{-1}(s),F^{-1}(t)\right)\\
&+\frac{1}{fF^{-1}(1-s)fF^{-1}(t)}K_F\left(F^{-1}(1-s),F^{-1}(t)\right)\\
&+\frac{1}{fF^{-1}(s)fF^{-1}(1-t)}K_F\left(F^{-1}(s),F^{-1}(1-t)\right)\\
&+\frac{1}{fF^{-1}(1-s)fF^{-1}(1-t)}K_F\left(F^{-1}(1-s),F^{-1}(1-t)\right)\Bigg\}.
\end{aligned}
$$

Corollary 5.1. *Suppose $X_1,\ldots,X_n$ are i.i.d. F, which is symmetric about μ and $\hat{F}$ is the corresponding empirical distribution function. We then have*

$$
K_G(s,t)=\frac{1}{4}-\frac{1}{2}f(\mu)\left\{\frac{s}{fF^{-1}(s)}+\frac{t}{fF^{-1}(t)}\right\}+\frac{1}{2}\frac{f^2(\mu)}{fF^{-1}(s)fF^{-1}(t)}(s\wedge t).
$$

Theorem 5.2. *Suppose $X_1,\ldots,X_n$ are observations from a distribution F. Let τ and $\hat{\tau}$ be as defined earlier. We then have*

$$
\sqrt{n}(\hat{\tau}-\tau)\xrightarrow{d}N(0,\ \omega^2)
$$

where

$$
\omega^2=\int_0^{1/2}\int_0^{1/2}K_G(s,t)dsdt.
$$

Corollary 5.2. *Suppose $X_1,\ldots,X_n$ are i.i.d. F which is symmetric about μ. Further, assume that F has finite second moment. We then have*

$$
\sqrt{n}\left(\hat{\tau}-\frac{1}{4}\right)\xrightarrow{d}N(0,\ \omega^2),
$$

where

$$
\omega^2=\frac{1}{16}-\frac{1}{4}f(\mu)E(|X-\mu|)+\frac{1}{4}f^2(\mu)E(X-\mu)^2. \tag{5.5}
$$

Table 5.1 Asymptotic variance of the symmetry statistic for some special symmetric distributions

Distribution	ω^2
Uniform	$\frac{1}{48} = 0.0208$
Normal	$\frac{\pi-2}{16\pi} = 0.0227$
Laplace	$\frac{1}{16} = 0.0625$
t_ν $(\nu > 2)$	$\frac{1}{16} - \left\{ \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \right\}^2 \frac{\nu-3}{4\pi(\nu-1)(\nu-2)}$
Beta(m, m)	$\frac{1}{16} - \frac{m+1}{16^m m(2m+1)} \left\{ \frac{\Gamma(2m)}{\Gamma^2(m)} \right\}^2$
Logistic	$\frac{12-24\log 2+\pi^2}{192} = 0.0273$

Note that the expression for ω^2 is invariant with respect to location and scale changes. Also note the similarity of the asymptotic variance formula to that of the asymptotic variance of S_K defined in [16].

Table 5.1 gives the asymptotic variance of the symmetry statistic for some special distributions. It is interesting to note that the asymptotic variance for the case of samples from t_ν distribution decreases with ν and $\to \frac{\pi-2}{16\pi}$ as $\nu \to \infty$. Similarly, for samples from the Beta(m, m) distribution, the asymptotic variance $\to \frac{\pi-2}{16\pi}$ as $m \to \infty$ and $\to \frac{1}{16}$ as $m \to 0$. The minimum is attained at $m = 1.66529$.

5.2.1 Estimation of the Asymptotic Variance

The asymptotic variance ω^2 of the $\hat{\tau}$ statistic given in (5.5) depends on unknown population quantities, which need to be consistently estimated. From [57] and [12], we see that the variance of $\frac{1}{n}\sum_{i=1}^n |X_i - \overline{X}|$ is of bounded order $O(n^{-1})$ and $E\frac{1}{n}\sum_{i=1}^n |X_i - \overline{X}| = E|X - \mu|\{1 + O(n^{-1/2})\}$. Hence, $\frac{1}{n}\sum_{i=1}^n |X_i - \overline{X}|$ is a consistent estimator of $E|X - \mu|$. Consequently, we can consistently estimate ω along the lines of [16] using

$$\hat{\omega}^2 = \frac{1}{16} - \frac{1}{4}\hat{f}(\hat{\mu})\frac{1}{n}\sum_{i=1}^n |X_i - \hat{\mu}| + \frac{1}{4}\hat{f}^2(\hat{\mu})S^2, \tag{5.6}$$

where $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ is the sample variance. Here $\hat{\mu}$ is the sample median and $\hat{f}(\hat{\mu})$ is the kernel density estimator of $f(\mu)$ obtained using

$$\hat{f}(x) = \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right),$$

where $K(\cdot)$ is the Epanechnikov kernel

$$K(t) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}t^2) & \text{for } |t| < \sqrt{5} \\ 0 & \text{otherwise,} \end{cases}$$

and the bandwidth h_n is chosen to be

$$h_n = 1.06 \min(\text{std. dev}, \text{IQR}/1.34) n^{-1/5}$$

(see [55, pp. 43–47] for more on these choices).

Note that for a given random sample $X_1, X_2, \ldots, X_n$, we can write

$$\hat{\tau} = \begin{cases} \frac{1}{n^2}\left[\#\left(X \le \frac{X_{(1)}+X_{(n)}}{2}\right) + \#\left(X \le \frac{X_{(2)}+X_{(n-1)}}{2}\right) + \cdots + \#\left(X \le \frac{X_{(n/2)}+X_{(n/2)+1}}{2}\right)\right] & \text{if } n \text{ is even} \\ \frac{1}{n^2}\left[\#\left(X \le \frac{X_{(1)}+X_{(n)}}{2}\right) + \#\left(X \le \frac{X_{(2)}+X_{(n-1)}}{2}\right) + \cdots + \#\left(X \le X_{(\frac{n+1}{2})}\right)\right] & \text{if } n \text{ is odd.} \end{cases}$$

Hence, a 2-sided test for symmetry at level of significance α can be given as: reject H_0 if $Z = \sqrt{n}\left|\frac{\hat{\tau}-\frac{1}{4}}{\hat{\omega}}\right| > Z_{\alpha/2}$.

5.3 Simulation Studies

We conducted a simulation study based on samples generated from a generalized lambda distribution (GLD) considered in [42, 58]. The quantile function of the generalized lambda distribution is given by (see [49])

$$x = F^{-1}(u) = \lambda_1 + \frac{u^{\lambda_3} - (1-u)^{\lambda_4}}{\lambda_2}, \qquad 0 \le u \le 1.$$

Here λ_1 is the location parameter of the distribution, λ_2 the scale parameter and λ_3, λ_4 determine the skewness, with symmetry when $\lambda_3 = \lambda_4$.

Data were generated from 14 members of the GLD family. For each $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ combination, we considered 4 sample sizes ($n = 30, 50, 100$ and 500). For each n, we generated a random sample of size n and performed four different tests of symmetry about an unknown location: (i) the proposed test (denoted by τ), (ii) the test proposed by [43] (denoted by T), (iii) the test proposed by [44] (denoted by γ_1) and (iv) the test proposed by [16] (denoted by C).

Note that each of the 4 test statistics are asymptotically normal whose asymptotic variances depend on the underlying distribution. In practice, one may decide to estimate this asymptotic variance along the lines of Sect. 5.2.1 to make these tests completely distribution free. However, a brief simulation study (not presented here) indicated that the resulting tests have poor behavior, especially for small samples. This is possibly due to the slow convergence of the estimator of the asymptotic variance and/or due to slow convergence to normality.

Consequently, we decided to follow the methodology used by [16, 43, 44], whereby the critical values of the 4 tests are first empirically estimated based on 10^5 simulations from a normal distribution. The resulting critical values are then used to study the power of the tests, with each power calculation based on 10^5 simulations.

Table 5.2 Estimated type I error probabilities of the various tests of symmetry when data are simulated from some symmetric members of the GLD family. Critical values are based on normal reference distribution. Results are based on 10^5 simulations, with nominal $\alpha = 0.05$

Test	P(type I error)			
	$n = 30$	$n = 50$	$n = 100$	$n = 500$
Case 1: $\lambda = (0, 2, 1, 1)$				
τ	0.0443	0.0495	0.0437	0.0411
T	0.0940	0.1029	0.1006	0.1064
γ_1	0.1151	0.1271	0.1308	0.1376
C	0.1173	0.1295	0.1314	0.1376
Case 2: $\lambda = (0, 0.19745, 0.134915, 0.134915)$				
τ	0.0539	0.0498	0.0507	0.0476
T	0.0531	0.0523	0.0518	0.0501
γ_1	0.0524	0.0491	0.0515	0.0496
C	0.0541	0.0505	0.0521	0.0496
Case 3: $\lambda = (0, -1, -0.08, -0.08)$				
τ	0.0928	0.0930	0.0938	0.0942
T	0.0609	0.0625	0.0635	0.0588
γ_1	0.0409	0.0446	0.0433	0.0382
C	0.0421	0.0457	0.0439	0.0381
Case 4: $\lambda = (0, -0.397912, -0.16, -0.16)$				
τ	0.1106	0.1229	0.1178	0.1344
T	0.0806	0.0827	0.0769	0.0773
γ_1	0.0459	0.0446	0.0444	0.0439
C	0.0476	0.0463	0.0449	0.0438
Case 5: $\lambda = (0, -1, -0.24, -0.24)$				
τ	0.1416	1475	0.1585	0.1820
T	0.1016	1069	0.1061	0.1074
γ_1	0.0457	0.0475	0.0454	0.0491
C	0.0478	0.0482	0.0461	0.0491
Case 6: $\lambda = (0, -5.11256, -1, -1)$				
τ	0.4300	0.4771	0.5413	0.6621
T	0.5364	0.5908	0.6663	0.7983
γ_1	0.0945	0.0864	0.0907	0.0991
C	0.0987	0.0895	0.0912	0.0989

Table 5.2 gives the results for the symmetric cases ($\lambda_1 = \lambda_2$). We see from Table 5.2 that for case 1 (uniform), the only recommended test is τ, since it is the only one maintaining the type I error level. For cases 3, 4 and 5, tests τ or T should not be used – instead one should preferably use γ_1. For case 6, none of the four tests are recommended. Finally, note that all the tests more or less maintain their significance levels for case 2. This is expected, since case 2 gives rise to a normal-like distribution.

Table 5.3 gives the results for the asymmetric cases. The density curves of the distributions are sketched in [42] and denoted there as case 2 to case 9. From Table 5.3, we see that for case 11, all the tests have very low power, even for quite large sample

Table 5.3 Estimated power of the various tests of symmetry when data are simulated from some asymmetric members of the GLD family. Critical values are based on normal reference distribution. Results are based on 10^5 simulations, with nominal $\alpha = 0.05$

Test	Power			
	$n = 30$	$n = 50$	$n = 100$	$n = 500$
Case 7: $\lambda = (0, 1, 1.4, 0.25)$				
τ	0.1872	0.2759	0.4627	0.9884
T	0.2692	0.3756	0.5668	0.9908
γ_1	0.2808	0.4002	0.5993	0.9933
C	0.2851	0.4041	0.5994	0.9933
Case 8: $\lambda = (0, 1, 0.00007, 0.1)$				
τ	0.6048	0.8141	0.9812	1.0000
T	0.6563	0.8388	0.9815	1.0000
γ_1	0.5972	0.8111	0.9787	1.0000
C	0.6032	0.8145	0.9788	1.0000
Case 9: $\lambda = (3.586508, 0.04306, 0.025213, 0.094029)$				
τ	0.2482	0.3850	0.6368	0.9993
T	0.2375	0.3533	0.5952	0.9985
γ_1	0.2019	0.3201	0.5690	0.9984
C	0.2071	0.3256	0.5696	0.9984
Case 10: $\lambda = (0, -1, -0.0075, -0.03)$				
τ	0.3981	0.5979	0.8678	1.0000
T	0.3793	0.5748	0.8424	1.0000
γ_1	0.3035	0.5001	0.8058	1.0000
C	0.3107	0.5063	0.8064	1.0000
Case 11: $\lambda = (-0.116734, 0.351663, -0.13, -0.16)$				
τ	0.1350	0.1489	0.2021	0.5145
T	0.0995	0.1076	0.1339	0.4132
γ_1	0.0565	0.0636	0.0879	0.3290
C	0.0587	0.0648	0.0885	0.3285
Case 12: $\lambda = (0, -1, -0.1, -0.18)$				
τ	0.2628	0.3813	0.6058	0.9974
T	0.2224	0.3279	0.5345	0.9927
γ_1	0.1309	0.2217	0.4249	0.9874
C	0.1360	0.2259	0.4255	0.9873
Case 13: $\lambda = (0, = -1, -0.001, -0.13)$				
τ	0.7935	0.9490	0.9989	1.0000
T	0.8398	0.9624	0.9986	1.0000
γ_1	0.7545	0.9370	0.9981	1.0000
C	0.7608	0.9390	0.9981	1.0000
Case 14: $\lambda = (0, -1, -0.0001, -0.17)$				
τ	0.8260	0.9616	0.9998	1.0000
T	0.8724	0.9715	0.9999	1.0000
γ_1	0.7908	0.9491	0.9997	1.0000
C	0.7960	0.9502	0.9997	1.0000

sizes. For cases 9, 10, 11 and 12, the test τ seems to be slightly preferable over the other choices. Note that in all these cases, the distribution is close to symmetric and is thus hard to distinguish. In none of the 8 cases, the test γ_1 seems to be preferable.

5.4 An Example

We use the data from [20] to illustrate our proposed method. Table II of [20] (reproduced in Table 5.4) gives survival times (in days) of 72 guinea pigs that received a dose of tubercle bacilli (see [11] for more details on the data). A plot of the kernel density estimate of the data is given in Fig. 5.1.

Table 5.4 Survival times (in days) of guinea pigs. Data reproduced from [20]

10	33	44	56	59	72	74	77	92	93
96	100	100	102	105	107	107	108	108	108
109	112	113	115	116	120	121	122	122	124
130	134	136	139	144	146	153	159	160	163
163	168	171	172	176	183	195	196	197	202
213	215	216	222	230	231	240	245	251	253
254	254	278	293	327	342	347	351	402	432
458	555								

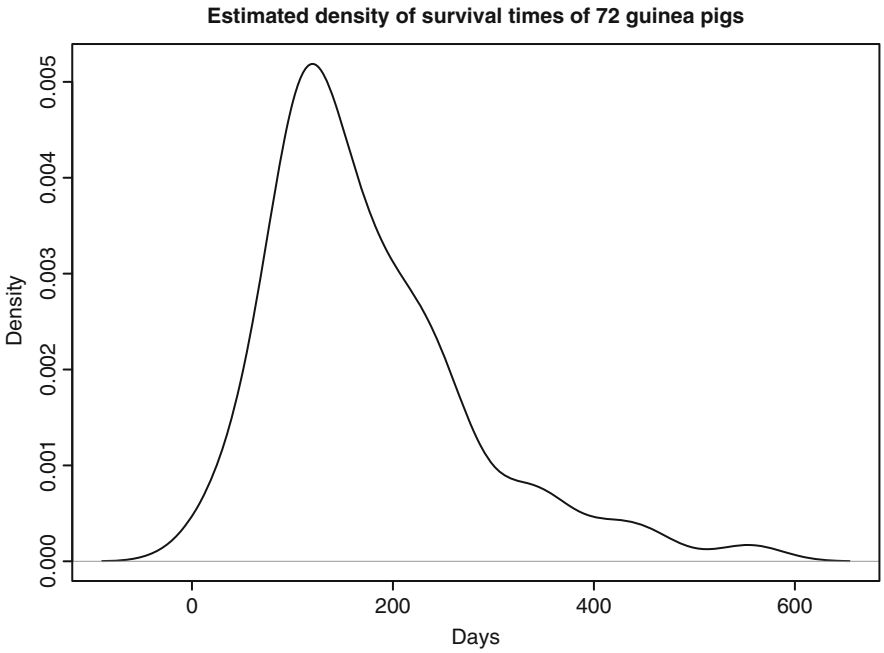


Fig. 5.1 Kernel density estimate of the guinea pig survival times

Application of the proposed test resulted in a test statistic value of $Z = 1.597538$, which corresponds to a p-value of 0.0550 against the alternative hypothesis of positive skewness. Thus, the data is close to significantly positively skewed at 5% level.

5.5 Conclusion

We have proposed a new test of univariate symmetry about an unknown location. From simulation studies, the test is seen to possess improved power over other choices when the underlying distribution is close to symmetric. The results presented here are based on estimating the critical values assuming a normal distribution. Preliminary simulations have shown that the asymptotically distribution-free version (obtained by estimating the asymptotic variance) has poor performance, especially for small sample sizes. We believe that this is due to the slow convergence to the limiting distribution. Methods to speed up convergence need further investigation.

Although we have presented the expression for the asymptotic variance only for the case of i.i.d sample, the same principle can be used for cases where one has ranked set samples or its generalizations (see [27]). The case of testing symmetry based on such generalized sampling schemes is currently ongoing and will be presented in a later article.

Appendix

Lemma 5.1. *Consider the map* $\phi : D[0,\ 1] \to D[0,\ 1]$ *given by*

$$\phi(F)(p) = F\left(\frac{F^{-1}(p) + F^{-1}(1-p)}{2}\right).$$

Let $\phi'_F(h)$ *be the Hadamard derivative of* $\phi(F)$ *in the direction of* h*. Then, we have*

$$\phi'_F(h)(p) = h\left(\frac{F^{-1}(p) + F^{-1}(1-p)}{2}\right) - \frac{1}{2} f\left(\frac{F^{-1}(p) + F^{-1}(1-p)}{2}\right) \times \left[\frac{h(F^{-1}(p))}{f(F^{-1}(p))} + \frac{h(F^{-1}(1-p))}{f(F^{-1}(1-p))}\right].$$

Proof of Lemma 5.1: The Hadamard derivative of $\phi(F)$ in the direction of h is given by

$$
\begin{aligned}
\phi'_F(h)(p) &= \lim_{t\to 0} \frac{\phi(F+th)(p)-\phi(F)(p)}{t} \\
&= \lim_{t\to 0} \frac{1}{t}[\phi(F+th)(p)-\phi(F)(p)] \\
&= \lim_{t\to 0} \frac{1}{t}\left[(F+th)\left(\frac{(F+th)^{-1}(p)+(F+th)^{-1}(1-p)}{2}\right)\right. \\
&\quad \left. -F\left(\frac{F^{-1}(p)+F^{-1}(1-p)}{2}\right)\right] \\
&= \lim_{t\to 0} \frac{1}{t}\left[F\left(\frac{(F+th)^{-1}(p)+(F+th)^{-1}(1-p)}{2}\right)\right. \\
&\quad +th\left(\frac{(f+th)^{-1}(p)+(F+th)^{-1}(1-p)}{2}\right) \\
&\quad \left. -F\left(\frac{F^{-1}(p)+F^{-1}(1-p)}{2}\right)\right] \\
&= \lim_{t\to 0} \frac{1}{t}\left[F\left(\frac{(F+th)^{-1}(p)+(F+th)^{-1}(1-p)}{2}\right)\right. \\
&\quad \left. -F\left(\frac{F^{-1}(p)+F^{-1}(1-p)}{2}\right)\right] \\
&\quad +h\left(\frac{(F+th)^{-1}(p)+(F+th)^{-1}(1-p)}{2}\right).
\end{aligned}
$$

Let $u = F^{-1}(p), u+v = (F+th)^{-1}(p), a = F^{-1}(1-p), a+b = (F+th)^{-1}(1-p)$. Then, $F(u) = p = F(u+v) + th(u+v)$ and $F(a) = 1-p = F(a+b) + th(a+b)$. Note that as $t \to 0$, $u \to 0$ and $v \to 0$. Hence,

$$
\begin{aligned}
h\left(\frac{(F+th)^{-1}(p)+(F+th)^{-1}(1-p)}{2}\right) &= h\left(\frac{u+v+a+b}{2}\right) \\
&\to h\left(\frac{u+a}{2}\right) \\
&= h\left(\frac{F^{-1}(p)+F^{-1}(1-p)}{2}\right).
\end{aligned} \tag{5.7}
$$

Also,

$$
\begin{aligned}
&\frac{1}{t}\left[F\left(\frac{(F+th)^{-1}(p)+(F+th)^{-1}(1-p)}{2}\right)-F\left(\frac{F^{-1}(p)+F^{-1}(1-p)}{2}\right)\right] \\
&= \frac{F\left(\frac{u+a}{2}+\frac{v+b}{2}\right)-F\left(\frac{u+a}{2}\right)}{\frac{v+b}{2}} \times \frac{v+b}{2t} \\
&= T_1 \times T_2 \text{ say.}
\end{aligned}
$$

As $t \to 0$, we have

$$\begin{aligned} T_1 &\to f\left(\frac{u+a}{2}\right) \\ &= f\left(\frac{F^{-1}(p)+F^{-1}(1-p)}{2}\right). \end{aligned} \tag{5.8}$$

Also, as $t \to 0$,

$$\begin{aligned} t_2 &= \frac{v+b}{2t} \\ &= \frac{1}{2}\left[\frac{vh(u+v)}{F(u)-F(u+v)} + \frac{bh(a+b)}{F(a)-F(a+b)}\right] \\ &\to \frac{1}{2}\left(\frac{-h(u)}{f(u)} + \frac{-h(a)}{f(a)}\right) \\ &= \frac{-1}{2}\left[\frac{h(F^{-1}(p))}{f(F^{-1}(p))} + \frac{h(F^{-1}(1-p))}{f(F^{-1}(1-p))}\right]. \end{aligned} \tag{5.9}$$

Combining (5.7), (5.8) and (5.9), we get the result. □

Proof of Theorem 5.1: From (5.2) and the functional delta method (see [4, 24]), we have

$$\sqrt{n}(\phi(\hat{F}) - \phi(F)) \xrightarrow{d} \phi'_F(W_F) \tag{5.10}$$

where

$$\begin{aligned} \phi'_F(W_F)(p) = {} & W_F\left(\frac{F^{-1}(p)+F^{-1}(1-p)}{2}\right) \\ & -\frac{1}{2} f\left(\frac{F^{-1}(p)+F^{-1}(1-p)}{2}\right) \\ & \times\left[\frac{W_F(F^{-1}(p))}{f(F^{-1}(p))} + \frac{W_F(F^{-1}(1-p))}{f(F^{-1}(1-p))}\right]. \end{aligned}$$

The result follows from the above. □

Proof of Theorem 5.2:

$$\sqrt{n}\left(\int_0^{1/2} \phi(\hat{F})(p)dp - \int_0^{1/2} \phi(F)(p)dp\right) \xrightarrow{d} \int_0^{1/2}\int_0^{1/2} K_G(s,t)dsdt.$$

□

Proof of Corollary 5.2: Note that

$$\int_0^{1/2}\int_0^{1/2} \frac{s}{fF^{-1}(s)} dsdt = \frac{1}{2}\int_{-\infty}^{\mu} (\mu - t) f(t) dt.$$

Also,

$$\int_0^{1/2}\int_0^{1/2} \frac{s \wedge t}{fF^{-1}(s) fF^{-1}(t)} dsdt = \int_{-\infty}^{\mu} (\mu - t)^2 f(t) dt.$$

Combining, we get the result. □

References

1. Ahmad IA (1996) Modification of some goodness of fit statistics II: Two-sample and symmetry testing. Sankhyá: Indian J Stat Ser A 58(3):464–472
2. Ahmad IA, Li Q (1997) Testing symmetry of an unknown density function by kernel method. J Nonparametr Stat 7(3):279–293
3. Aki S (1987) On nonparametric tests for symmetry. Ann Inst Stat Math 39:457–472
4. Andersen PK, Borgan O, Gill RD, Keiding N (1993) Statistical methods based on counting processes. Springer, New York
5. Antille A, Kersting G (1977) Tests for symmetry. Z Wahrsch Verw Gebiete 39:235–255
6. Antille A, Kersting G, Zucchini W (1982) Testing symmetry. J Am Stat Assoc 77(379): 639–646
7. Arcones MA, Gine E (1991) Some bootstrap tests for symmetry for univariate continuous distributions. Ann Stat 19(3):1496–1511
8. Baklizi A (2003) A conditional distribution free runs test for symmetry. J Nonparametr Stat 15(6):713–718
9. Baringhaus L, Henze N (1992) A characterization of new and consistent tests for symmetry. Commun Stat Theory Methods 21(6):1555–1566
10. Bhattacharya PK, Gastwirth JL, Wright AL (1982) Two modified Wilcoxon tests for symmetry about an unknown location parameter. Biometrika 69(2):377–382
11. Bjerkedal T (1960) Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli. Am J Hyg 72:130–148
12. Bonett DG, Seier E (2003) Confidence intervals for mean absolute deviations. Am Statistcian 57(4):233–236
13. Boos DD (1982) A test for asymmetry associated with the Hodges–Lehmann estimator. J Am Stat Assoc 77(379):647–651
14. Butler CC (1969) A test for symmetry using the sample distribution function. Ann Math Stat 40(6):2209–2210
15. Cabana A, Cabana EM (2000) Tests of symmetry based on transformed empirical processes. Can J Stat 28(4):829–839
16. Cabilio P, Masaro J (1996) A simple test of symmetry about an unknown median. Can J Stat 24(3):349–361
17. Cohen JP, Menjoge SS (1988) One-sample run tests of symmetry. J Stat Plan Inference 18:93–100
18. CsörgHo S, Heathcote CR (1987) Testing for symmetry. Biometrika 74(1):177–184
19. Davis CE, Quade D (1978) U-statistics for skewness and symmetry. Commun Stat Theory Methods 7(5):413–418
20. Doksum KA (1975) Measures of location and asymmetry. Scand J Stat 2(1):11–22
21. Doksum KA, Fenstad G, Aaberge R (1977) Plots and tests for symmetry. Biometrika 64(3):473–487

22. Ekström M, Jammalamadaka SR (2007) An asymptotically distribution-free test of symmetry. J Stat Plan Inference 137:799–810
23. Eubank RL, Iariccia VN, Rosenstein RB (1992) Testing symmetry about and unknown median, via linear rank procedures. J Nonparametr Stat 1(4):301–311
24. Fernholz LT (1983) Von Mises calculus for statistical functionals. Springer, New York
25. Finch SJ (1977) Robust univariate test of symmetry. J Am Stat Assoc 72(358):387–392
26. Gastwirth JL (1971) On the sign test for symmetry. J Am Stat Assoc 66(336):821–823
27. Ghosh K, Tiwari RC (2009) A unified approach to variations of ranked set sampling with applications. J Nonparametr Stat 21(4):471–485
28. Gupta MK (1967) An asymptotically nonparametric test of symmetry. Ann Math Stat 38(3):849–866
29. Hájek J (1962) Asymptotically most powerful rank-order tests. Ann Math Stat 33:1124–1147
30. Hájek J, Sidák Z (1967) Theory of rank tests. Academic, New York
31. Henze N (1993) On the consistency of a test for symmetry based on a runs statistic. J Nonparametr Stat 3(2):195–199
32. Hill DL, Rao PV (1977) Tests of symmetry based on Cramer–von Mises statistics. Biometrika 64(3):489–494
33. Hill DL, Rao PV (1981) Tests of symmetry based on Watson statistics. Commun Stat Theory Methods 10(11):1111–1125
34. Hotelling H, Solomons LM (1932) The limits of a measure of skewness. Ann Math Stat 3(2):141–142
35. Huber PJ (1972) Robust statistics: A review. Ann Math Stat 43:1041–1067
36. John JA, Draper NR (1980) An alternative family of transformations. Appl Stat 29(2):190–197
37. Koopman PAR (1979) Testing symmetry with a procedure combining the sign test and the sign rank test. Stat Neerl 33:137–141
38. Koutrouvelis IA (1985) Distribution-free procedures for location and symmetry inference problems based on the empirical characteristic function. Scand J Stat 12(4):257–269
39. Koziol JA (1980) On a Cramér–von Mises-type statistic for testing symmetry. J Am Stat Assoc 75(369):161–167
40. Koziol JA (1985) A note on testing symmetry with estimated parameters. Stat Probab Lett 3:227–230
41. Lockhart RA, McLaren CG (1985) Asymptotic points for a test of symmetry about a specified median. Biometrika 72(1):208–210
42. McWilliams TP (1990) A distribution-free test for symmetry based on a runs statistic. J Am Stat Assoc 85(412):1130–1133
43. Miao W, Gel YR, Gastwirth JL (2006) A new test of symmetry about an unknown median. In: Random walk, sequential analysis and related topics: A festschrift in honor of Yuan-Shih Chow. World Scientific, New Jersey, pp 199–214
44. Mira A (1999) Distribution-free test for symmetry based on Bonferroni's measure. J Appl Stat 26(8):959–972
45. Modarres R, Gastwirth JL (1996) A modified runs test for symmetry. Stat Probab Lett 31: 107–112
46. Modarres R, Gastwirth JL (1998) Hybrid test for the hypothesis of symmetry. J Appl Stat 25(6):777–783
47. Orlov AI (1972) On testing the symmetry of distributions. Theory Prob Appl 17:357–361
48. Psaradakis Z (2003) A bootstrap test for symmetry of dependent data based on a Kolmogorov–Smirnov type statistic. Comm Stat Simul Comput 32(1):113–126
49. Ramberg JS, Schmeiser BW (1974) An approximate method for generating asymmetric random variables. Commun ACM 17:78–82
50. Randles RH, Fligner MA, Policello II GE, Wolfe DA (1980) An asymptotically distribution-free test for symmetry versus asymmetry. J Am Stat Assoc 75(369):168–172
51. Reynolds MR Jr (1975) A sequential signed-rank test for symmetry. Ann Stat 3(2):382–400
52. Rosenstein RB (1989) Tests of symmetry derived as components of Pearson's phi-squared distance measure. Commun Stat Theory Methods 18(5):1617–1626

53. Rothman ED, Woodroofe M (1972) A Cramér von–Mises type statistic for testing symmetry. Ann Math Stat 43(6):2035–2038
54. Schuster EF, Barker RC (1987) Using the bootstrap in testing symmetry versus asymmetry. Commun Stat Simul Comput 16(1):69–84
55. Silverman BW (1986) Density estimation for statistics and data analysis. Chapman and Hall, London
56. Srinivasan R, Godio LB (1974) A Cramer–von Mises type statistic for testing symmetry. Biometrika 61(1):196–198
57. Stuart A, Ord JK (1994) Kendall's advanced theory of statistics, vol 1, 6 edn. Halsted, New York
58. Tajuddin IH (1994) Disribution-free test for symmetry based on Wilcoxon two-sample test. J Appl Stat 21(5):409–415
59. van Eeden C, Bernard A (1957) A general class of distribution-free tests for symmetry containing the tests of Wilcoxon and Fisher. Nederl Akad Wetensch A60:381–408

Chapter 6
A Semiparametric Bayesian Method of Clustering Genes Using Time-Series of Expression Profiles

Arvind K. Jammalamadaka and Kaushik Ghosh

Abstract An increasing number of microarray experiments look at expression levels of genes over the course of several points in time. In this article, we present two models for clustering such time series of expression profiles. We use nonparametric Bayesian methods which make the models robust to misspecifications and provide a natural framework for clustering of the genes through the use of Dirichlet process priors. Unlike other clustering techniques, the resulting number of clusters is completely data driven. We demonstrate the effectiveness of our methodology using simulation studies with artificial data as well as through an application to a real data set.

6.1 Introduction

Rapid advances in genome-sequencing technology has made it possible to collect data simultaneously on expression-levels of all genes of an organism relatively easily than it has been ever before. Usually, the number of genes is in the thousands and the data might be very high-dimensional, thus requiring specialized statistical tools for analysis. One of the natural questions that scientists have been pondering over is how to use such expression data to cluster genes into functionally related groups. The aim is thus to identify genes with similar expression patterns, which can further lead to improved understanding of cell functioning – possibly opening way for better treatment of diseases.

The expression data can be in the form of a single snapshot at a point of time or can be in the form of multiple measurements of the genes, taken at several points in time. The latter gives rise to time series of expression profiles and is naturally more informative in clustering genes than if one were to use single time snapshots.

A.K. Jammalamadaka (✉)
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
e-mail: ajamma@gmail.com

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_6,

Time series experiments are becoming increasingly common (see [2,8–10,12,13,15] for some recent examples of work in this area) with the improvements in high-throughput data collection techniques such as microarrays.

Analyzing such expression data present several special challenges, since one needs to account for the correlation among several genes at a timepoint as well as correlation among measurements on the same gene at different timepoints. Several of the past methods dealing with such data have assumed that the expression profile for each gene is a multivariate vector where the profile vector, given the parameters, is assumed to be multivariate normal with independent components (see, e.g., [8–10]). While such simpler models provide a first-cut, they fail to take into account the dependence of the measurements on a gene at various timepoints and are thus not realistic. In this paper, we in contrast, allow for potentially rich correlation structures between timepoints. Such dependence between expression values for the same gene, is quite natural in our opinion, and is the main contribution of this paper.

This paper is organized as follows. In Sect. 6.2, we present two proposed models for the time series of expression data. In Sect. 6.3, we discuss the techniques used to estimate the model parameters. To demonstrate the ability of the proposed models to correctly identify the underlying clusters, we present results of a short simulation study in Sect. 6.4. In Sect. 6.5, we illustrate this method by an application to the well-studied yeast cell-cycle data [14]. Finally, we present concluding remarks in Sect. 6.6.

6.2 Model Description

Suppose that a microarray experiment consists of N genes. Each gene is repeatedly measured over time on p different occasions, resulting in a p-dimensional measurement vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})'$ for gene i ($i = 1, \ldots, N$). Let $\boldsymbol{\mu}_i$ be the mean vector of $\boldsymbol{y}_i$ and $\boldsymbol{\Sigma}_i$ be the corresponding covariance matrix. In what follows, we present two different models for the joint distribution of the measurement vectors $\boldsymbol{y}_i$. The models differ in how the variance-covariance matrix $\boldsymbol{\Sigma}_i$ is represented.

6.2.1 Heteroscedastic Model

In this model, we assume that the variance-covariance matrix for the measurements on gene i is given by $\boldsymbol{\Sigma}_i = \sigma_i^2 H(\rho_i)$. Here σ_i^2 and ρ_i are gene-specific variance and one-step correlation parameters respectively and $H(\cdot)$ is a matrix-variate function that is common to all genes. The matrix $H(\cdot)$ provides the dependence structure among the measurements at various timepoints on a particular gene and is assumed to be completely known.

Two possible forms of the $H(\cdot)$ matrix are

$$H(\rho) = ((\rho^{|i-j|}))_{p\times p}$$

and

$$H(\rho) = ((I_{\{i=j\}} + \rho I_{\{i\neq j\}}))_{p\times p}.$$

The first form assumes that the correlation between measurements at any two timepoints decays exponentially as a function of the distance between the two timepoints, so points further away are less correlated with one another. The second form assumes that the correlation remains the same between any two points of time.

We assume that conditional on the mean vector, the common variance and one-step correlation, gene expression profiles for various genes are independent multivariate normal random vectors. That is,

$$\boldsymbol{y}_i|(\boldsymbol{\mu}_i, \sigma_i^2,\ \rho_i) \overset{indep}{\sim} N_p(\boldsymbol{\mu}_i,\ \sigma_i^2 H(\rho_i)), \qquad i = 1, \ldots, N.$$

Let $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i,\ \sigma_i^2,\ \rho_i)$ be the vector of parameters used to describe the joint distribution of the expression profile of ith gene. We will say two genes i and j cluster together if their corresponding parameter vectors $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ are identical. To this end, we assume that $\boldsymbol{\theta}_i|G \overset{iid}{\sim} G$ and $G|(M, G_0) \sim \mathcal{D}(M,\ G_0)$. Here $\mathcal{D}(M,\ G_0)$ denotes a Dirichlet process [5] with base measure G_0 and concentration M.

The reasons for such a choice of prior distribution are 2-fold. First, using a Dirichlet process makes the distribution of the parameter vectors flexible and robust to mis-specifications. Second, due to the natural clustering property of samples from a distribution with a Dirichlet process prior [1], there is inherent data-driven clustering among the model parameters.

We assume that under the baseline prior G_0, $(\boldsymbol{\mu}_i, \sigma_i^2, \rho_i)$ are independent with multivariate normal, inverted-gamma and beta distributions, respectively. In other words,

$$G_0 = N_p(\mu_0, \Sigma_0) \times IG(a_\sigma, b_\sigma) \times Beta(a_\rho, b_\rho)$$

where

$$\Sigma_0 = diag(\sigma_{01}^2, \ldots, \sigma_{0p}^2)$$

with

$$\sigma_{01}^2, \ldots, \sigma_{0p}^2 \overset{iid}{\sim} IG(a_0, b_0)$$

and

$$\boldsymbol{\mu}_0 = (\mu_{01}, \ldots, \mu_{0p})^T.$$

We further assume

$$M \sim Gamma(a_M, b_M).$$

All the hyperparameters $a_\sigma, b_\sigma, a_0, b_0, a_\rho, b_\rho,\ a_M, b_M, \mu_{01}, \ldots, \mu_{0p}$ are assumed known and in practice, are chosen to make the priors as noninformative as possible. Due to lack of closed form expressions for the posterior distribution

and easy availability of univariate conditionals, we will use a Markov chain Monte Carlo approach to estimate the model parameters. The natural clustering property of the Dirichlet Process will lead to clustering of the values of $\boldsymbol{\theta}_i$.

6.2.2 Homoscedastic Model

In this case, we assume the variance of measurements is same across all genes and across all timepoints. Hence, the clustering of genes is achieved based on the gene-specific mean vector and one-step correlation. Thus, we have

$$\boldsymbol{y}_i|(\boldsymbol{\mu}_i,\ \sigma^2,\ \rho_i) \overset{indep}{\sim} N_p(\boldsymbol{\mu}_i, \sigma^2 H(\rho_i)), \quad i = 1, \ldots, N,$$

where N = number of genes,

$$H(\rho) = ((\rho^{|i-j|}))_{p\times p},$$

or

$$H(\rho) = ((I_{\{i=j\}} + \rho I_{\{i\neq j\}}))_{p\times p}.$$

is the correlation matrix, and p = number of observations per gene.

We further assume:

$$\begin{aligned}
(\boldsymbol{\mu}_i, \rho_i) \equiv \theta_i &\overset{iid}{\sim} G,\\
G &\sim DP(M, G_0),\\
G_0 &= N_p(\boldsymbol{\mu}_0, \Sigma_0) \times Beta(a_\rho, b_\rho),\\
\Sigma_0 &= diag(\sigma_1^2, \cdots, \sigma_p^2)_{p\times p},\\
\sigma_1^2, \ldots, \sigma_p^2 &\overset{iid}{\sim} IG(a^*, b^*),\\
\boldsymbol{\mu}_0 &= (\mu_{01}, \ldots, \mu_{0p})',\\
M &\sim Gamma(a_M, b_M),\\
\sigma^2 &\sim IG(a_\sigma, b_\sigma).
\end{aligned}$$

Note that this model differs from the heteroscedastic model by the fact that all genes are assumed to have the same variability in their pointwise measurements and cannot be obtained as a special case of the former model.

6.3 Model Fitting and Estimation

Due to lack of closed-form expressions of the posterior distributions, model fitting is done using Markov chain Monte Carlo procedures. In particular, since univariate conditionals of the model parameters are easily obtained, we use the Gibbs sampler. Details of the implementation of the Gibbs sampler are outlined below.

6.3.1 Heteroscedastic Model

We can update the model parameters according the following scheme:

1. $$\sigma_{0j}^2|\cdots \sim IG\left(a_0+\frac{N}{2},\left[\frac{1}{b_0}+\frac{1}{2}\sum_{i=1}^{N}(\mu_{ij}-\mu_{0j})^2\right]^{-1}\right),$$
   $j=1,\ldots,p$.
2. Due to the clustering property of samples generated from Dirichlet process, not all values of θ_i are distinct. In any iteration, let there be k distinct values, $\theta_1^*,\ldots,\theta_k^*$. Let $\boldsymbol{c}=(c_1,\ldots,c_N)$ be the configuration vector where $c_i=j$ if and only if $\theta_i=\theta_j^*$, $j=1,\ldots,k$. Let n_j be the number of occurrences of θ_j^*. The configuration vector is updated every iteration using algorithms of [11].
3. Update the distinct values θ_j^* as follows ($j=1,\ldots,k$)

$$f(\mu_j^*,{\sigma_j^*}^2,\rho_j^*|\cdots)\propto\prod_{i:c_i=j}f(y_i|\mu_j^*,{\sigma_j^*}^2,\rho_j^*)f(\mu_j^*|\mu_0,\Sigma_0)\times f({\sigma_j^*}^2|a_\sigma,b_\sigma)f(\rho_j^*|a_\rho,b_\rho).$$

Consequently, it is easy to show that

(a) $\mu_j^*|\cdots\sim N(\eta_j^*,\Gamma_j^*)$ where

$$\Gamma_j^*=\left[\left(\frac{{\sigma_j^*}^2}{n_j}H(\rho_j^*)\right)^{-1}+\Sigma_0^{-1}\right]^{-1}$$

and

$$\eta_j^*=\Gamma_j^*\left[\left(\frac{{\sigma_j^*}^2}{n_j}H(\rho_j^*)\right)^{-1}\left(\frac{1}{n_j}\sum_{i:c_i=j}y_i\right)+\Sigma_0^{-1}\mu_0\right]$$

(b)

$${\sigma_j^*}^2|\cdots\sim IG\left(a_\sigma+\frac{pn_j}{2},\left[\frac{1}{b_\sigma}+\frac{1}{2}\sum_{i:c_i=j}(y_i-\mu_j^*)^TH(\rho_j^*)^{-1}(y_i-\mu_j^*)\right]^{-1}\right),$$

(c)

$$f(\rho_j^*|\cdots)\propto\frac{1}{|H(\rho_j^*)|^{\frac{n_j}{2}}}e^{-\left[\frac{1}{2{\sigma_j^*}^2}\sum_{i:c_i=j}(y_i-\mu_j^*)^TH(\rho_j^*)^{-1}(y_i-\mu_j^*)\right]}\times(\rho_j^*)^{a_\rho-1}(1-\rho_j^*)^{b_\rho-1}. \tag{6.1}$$

The components of θ_j^* can be updated one at a time using the above results. Updating of ρ_j^* using (6.1) can be done using Adaptive Rejection Sampling (ARS) or Adaptive Rejection Metropolis Sampling (ARMS) (see [6, 7]).

4. M is updated using a mixture of Gammas as outlined in [4].

6.3.2 Homoscedastic Model

The parameter updates are done according to the following scheme:

1. Update the pointwise variance using

$$\sigma^2|\cdots \sim IG\left(a_\sigma + \frac{np}{2}, \left[\frac{1}{b_\sigma} + \frac{1}{2}\sum_{i=1}^{N}(y_i - \mu_i)H(\rho_i)^{-1}(y_i - \mu_i)\right]^{-1}\right).$$

2. Update the entries of the baseline covariance matrix Σ_0 using

$$\sigma_j^2|\cdots \sim IG\left(a^* + \frac{N}{2}, \left[\frac{1}{b^*} + \frac{1}{2}\sum_{i=1}^{N}(\mu_{ij} - \mu_{0j})^2\right]^{-1}\right),$$

   $j = 1, \ldots, p$.

3. As in the heteroscedastic model, let there be k distinct values, $\theta_1^*, \ldots, \theta_k^*$ and let $\boldsymbol{c} = (c_1, \ldots, c_N)$ be the configuration vector where $c_i = j$ if and only if $\theta_i = \theta_j^*$, $j = 1, \ldots, k$. Let n_j be the number of occurrences of θ_j^*. We update the configuration vector using algorithms of [11].
4. Update the distinct values as follows:

$$\boldsymbol{\mu}_j^*|\cdots \sim N_p(\boldsymbol{\mu}_j^*, \Gamma_j^*),$$

   where

$$\Gamma_j^* = \left[\left(\frac{\sigma^2}{n_j}H(\rho_j)\right)^{-1} + \Sigma_0^{-1}\right]^{-1}$$

   and

$$\mu_j^* = \Gamma_j^*\left[\left(\frac{\sigma^2}{n_j}H(\rho_j)\right)^{-1}\left(\frac{1}{n_j}\sum_{i:c_i=j} y_i\right) + \Sigma_0^{-1}\mu_0\right]$$

$$f(\rho_j^*|\cdots) \propto \frac{1}{|H(\rho_j^*)|^{\frac{n_j}{2}}} \times e^{-\left[\frac{1}{2\sigma^2}\sum_{i:c_i=j}(y_i-\mu_j^*)^T H(\rho_j^*)^{-1}(y_i-\mu_j^*)\right]}(\rho_j^*)^{a_\rho-1}(1-\rho_j^*)^{b_\rho-1}.$$

Use ARS or ARMS to do the last update.

5. Update M using a mixture of Gammas, along the lines of [4].

6.4 A Simulation Study

Performance of both the heteroscedastic and homoscedastic models were evaluated using a simulation study involving a synthetic data set. This data consists of expression on $N = 60$ genes measured at $p = 10$ equally-spaced timepoints, without loss of generality which are taken to be $t_1 = 1, \ldots, t_{10} = 10$. The 60 genes are divided into 3 groups, each consisting of 20 genes. The data were generated according to the following model

$$y_{ij} = \psi_i(t_j) + e_{ij}$$

where

$$\psi_i(t) = \delta_{i0} + \delta_{i1}t + \beta_{i1}(t-t_1)_+ + \cdots + \beta_{ip}(t-t_p)_+$$

is the mean trajectory of the ith group and e_{ij} is the random error in measuring the expression level for gene i at timepoint t_j. The parameters for the mean trajectories for the three groups are given in Table 6.1. The errors for each group are assumed to be normal with zero mean and variances given by 0.25, 1.0 and 0.04 respectively. In addition, we also assumed that there is a correlation of $\rho = 0.3$ between time-points.

Once we generated the data, the Gibbs sampler was used to sample from the posterior distribution for the fitted models. Both the homoscedastic and heteroscedastic models were fitted to the simulated data. A burn-in of 10,000 iterations was found to be enough to reach steady-state. An additional 10,000 iterations were used for the

Table 6.1 Synthetic data trajectory parameters

	Group 1	Group 2	Group 3
δ_{i0}	0	0	−1.0
δ_{i1}	0.5	0.5	0.2
β_{i1}	0	0	0
β_{i2}	0	0	0
β_{i3}	0	0	0
β_{i4}	0	0	0
β_{i5}	0	−1.5	0
β_{i6}	0	0	0
β_{i7}	0	0	0
β_{i8}	0	0	0

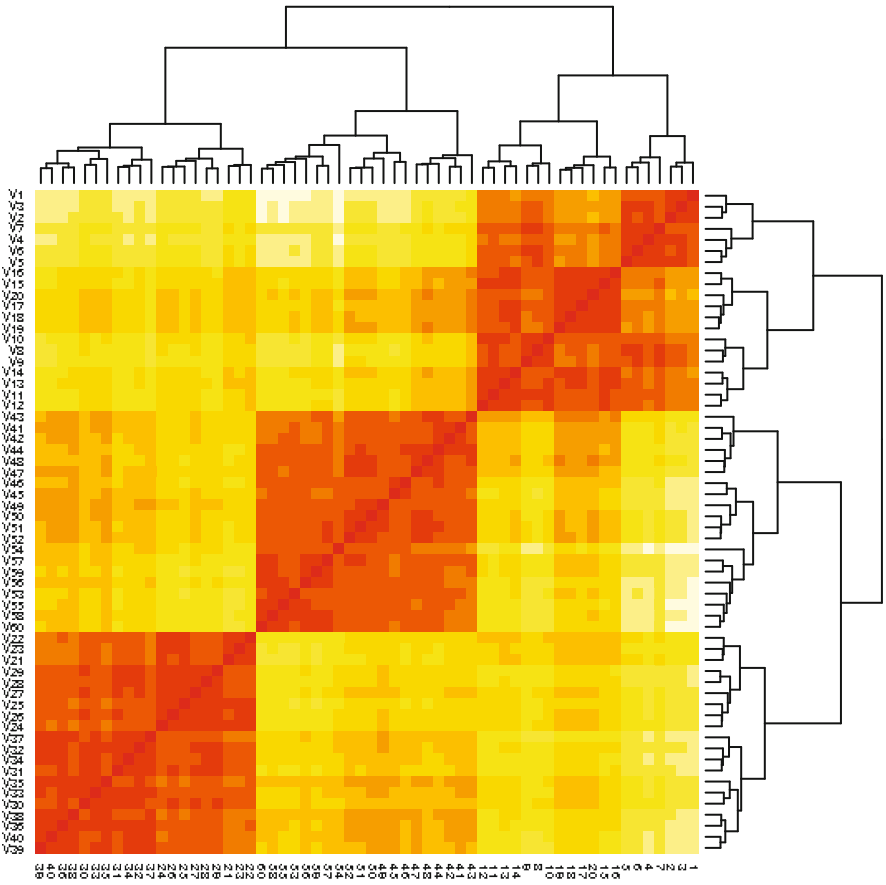


Fig. 6.1 Heatmap of clusters of the synthetic data based on the average incidence matrix from the heteroscedastic model. Higher intensity (in *deeper red*) denotes more likelihood of clustering together, while lower intensity (in *lighter yellow*) denotes less likelihood of clustering. Note that the three original groups of 20 units each are fully recovered (denoted by the three large squares on the diagonal)

purposes of posterior inference. For each iteration l, we obtained a configuration vector $c^{(l)} = (c_1^{(l)}, \ldots, c_N^{(l)})$ of the cluster-structure of the genes. This gave rise to an incidence matrix $Q_{N\times N}^{(l)}$, where $Q^{(l)} = ((q_{i,j}^{(l)}))$ and $q_{i,j}^{(l)} = 1$ if $c_i^{(l)} = c_j^{(l)}, 0$ otherwise. The average incidence matrix $\overline{Q} = \frac{1}{L}\sum_{l=1}^{10,000} Q^{(l)}$ was then computed. The `heatmap` routine in `R` was used on $1 - \overline{Q}$ to generate a heatmap of the clustering of the genes based on this average incidence matrix. The result for the heteroscedastic model is shown in Fig. 6.1.

We note that these models generally seem to perform very similarly on the test data. While both of them are able to isolate members of the same original group into clusters, the clustering is not perfect in all the iterations.

6.5 Application to Spellman Data

The proposed models were next applied to a standard subset of the well-known data on time series of gene expression of yeast cells due to [14]. Details on the data and how to access it can be obtained from http://genome-www.stanford.edu/cellcycle. The subset consists of observations on $N = 798$ genes measured at $p = 18$ time-points. Both the homoscedastic and the heteroscedastic models were fit to the data. A burn-in of 10,000 iterations stabilized the sampler. After that, an additional 5,000 iterations were used to estimate the model parameters. Each iteration produced a configuration structure of the genes that cluster together, giving rise to an incidence matrix. The average incidence matrix was used to ascertain the "average cluster structure" of the genes using the method of [3] or through heatmaps as described earlier. The resulting heatmap for the heteroscedastic model is provided in Fig. 6.2.

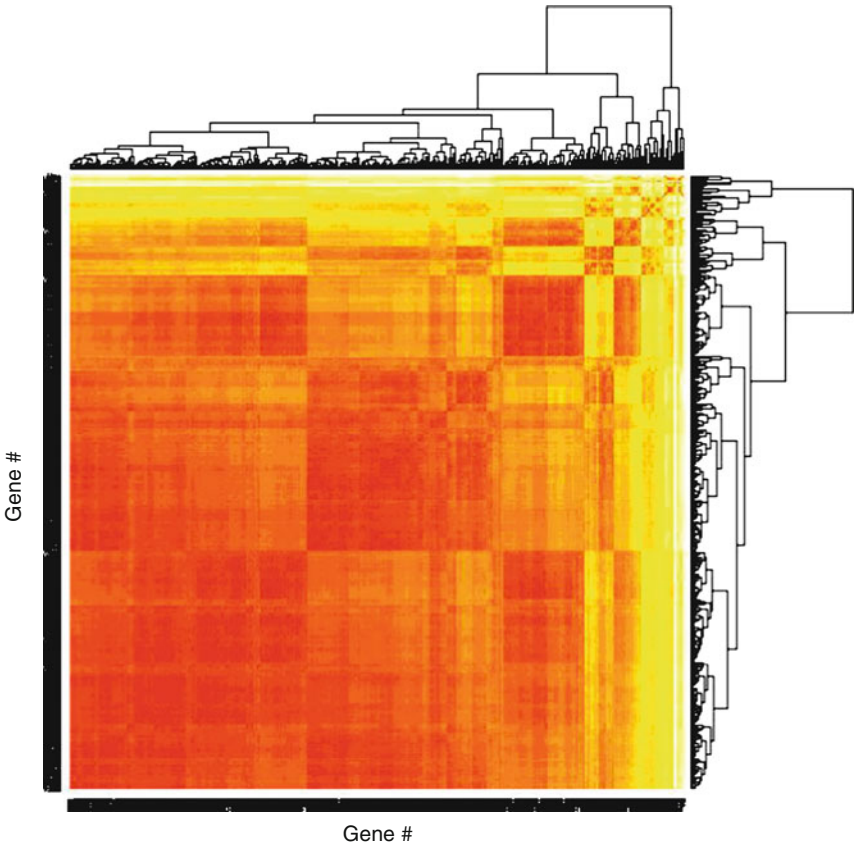


Fig. 6.2 Heatmap of clusters of the 798 genes in the Spellman data based on the average incidence matrix for the heteroscedastic model. Higher intensity of *red* denotes more likelihood of clustering together

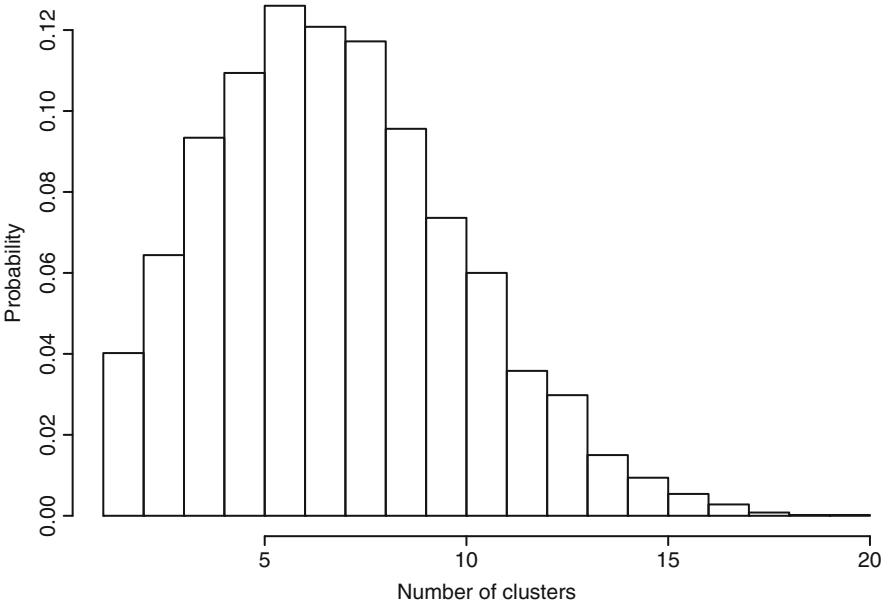


Fig. 6.3 Histogram of number of clusters generated using the heteroscedastic model on the Spellman data

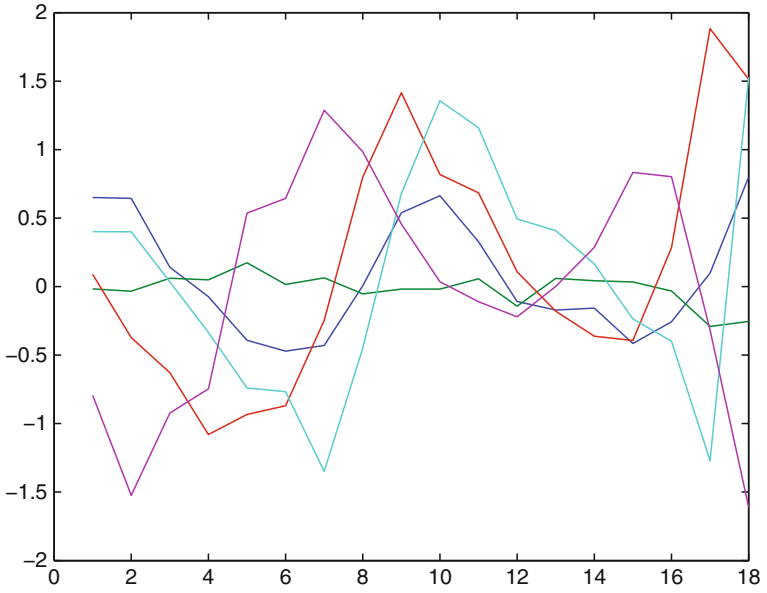


Fig. 6.4 Mean curves of top 5 clusters found by applying the heteroscedastic model to the Spellman data

The median number of clusters using the homoscedastic model is 5 with a 95% credible interval being (1, 13) and the corresponding figures for the heteroscedastic model were 7 and (2, 14) respectively. The histogram of the number of clusters in the heteroscedastic model is given in Fig. 6.3. All the above models were coded in C, and the code can be obtained from the authors upon request.

A plot of the mean expression profiles for the top 5 clusters corresponding to the heatmap and the resulting dendrogram for the heteroscedastic model is given in Fig. 6.4. To get a rough measure of cluster separation, we consider a rather crude measure of distance achieved by each of the (cluster-specific) average trajectories from the overall mean trajectory $\overline{y}$, as measured by the sum across the 5 clusters and the 18 timepoints using

$$\sum_{i=1}^{5}\sum_{t=1}^{18}(y_{it}-\overline{y}_t)^2.$$

This measure has a value of 39.68 for the heteroscedastic model and 40.13 for the homoscedastic model. Hence, in terms of the measure of separation, of the clusters the two models are quite close.

As a check of how robust our choice of prior parameters on M are, we changed these to yield an expected number of clusters that is closer to 5. The resulting mean curves for the top 5 clusters are nearly indistinguishable from those obtained previously. This assures us that the hyperparameters do not play a very significant role, which is often a point of concern in Bayes methods.

6.6 Conclusion

We have provided a semiparametric Bayesian approach to clustering genes based on their expression profiles, which we assume have certain correlation structure. This method is easily implemented using standard computer languages and produces results in a reasonable period of time (it took about 2 hours to cluster the 798 genes on an Apple MacBook Pro with 2.6 GHz Intel Core 2 Duo and 4 GB of RAM). One can easily add additional steps to the Gibbs sampler to make the model more robust to mis-specifications (e.g., using a multivariate t instead of multivariate normal). The model can also be used to handle missing data by replacing them with their imputed values or as an additional MCMC step.

References

1. Antoniak CE (1974) Mixtures of Dirichlet processes with applications to nonparametric problems. Ann Stat 2:1152–1174
2. Bar-Joseph Z, Gerber G, Jaakkola T, Gifford D, Simon I (2003) Continuous representations of time series gene expression data. J Comput Biol 3:341–356

3. Dahl D (2006) Model-based clustering for expression data via a Dirichlet process mixture model. In: Bayesian inference for gene expression and proteomics. Cambridge University Press, Cambridge, pp 201–218
4. Escobar M, West M (1995) Bayesian density estimation and inference using mixtures. J Am Stat Assoc 90(430):577–588
5. Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 1(2): 209–230
6. Gilks WR, Best NG, Tan KKC (1995) Adaptive rejection metropolis sampling. Appl Stat 44: 455–472
7. Gilks WR, Wild P (1992) Adaptive rejection sampling for Gibbs sampling. Appl Stat 41(2): 337–348
8. Liu X, Sivaganesan S, Yeung K, Guo J, Baumgarner RE, Medvedovic M (2006) Context-specific infinite mixtures for clustering gene expression profiles across diverse microarray data set. Bioinformatics 22:1737–1744
9. Medvedovic M, Sivaganesan S (2002) Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics 18(9):1194–1206
10. Medvedovic M, Yeung KY, Baumgarner R (2004) Bayesian mixture model based clustering of replicated microarray data. Bioinformatics 20:1222–1232
11. Neal RM (2000) Markov chain sampling methods for Dirichlet process mixture models. J Comput Graph Stat 9(2):249–265
12. Qian J, Dolled-Filhart M, Lin J, Yu H, Gerstein M (2001) Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. J Mol Biol 314:1053–1066
13. Singh R, Palmer N, Gifford D, Berger B, Bar-Joseph Z (2005) Active learning for sampling in time-series experiments with application to gene expression analysis. In: ICML '05: proceedings of the 22nd international conference on Machine learning. ACM, New York, pp 832–839
14. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Botstein D, Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cervisiae by microarray hybridization. Mol Biol Cell 9(12):3273–3297
15. Yuan M, Kendziorski C (2006) Hidden Markov models for microarray time course data in multiple biological conditions. J Am Stat Assoc 101(476):1323–1332

Chapter 7
On Implementation of the Markov Chain Monte Carlo Stochastic Approximation Algorithm

Yihua Jiang, Peter Karcher, and Yuedong Wang

Abstract The Markov Chain Monte Carlo Stochastic Approximation Algorithm (MCMCSAA) was developed to compute estimates of parameters with incomplete data. In theory this algorithm guarantees convergence to the expected fixed points. However, due to its flexibility and complexity, care needs to be taken for implementation in practice. In this paper we show that the performance of MCMCSAA depends on many factors such as the Markov chain Monte Carlo sample size, the step-size of the parameter update, the initial values and the choice of an approximation to the Hessian matrix. Good choices of these factors are crucial to the practical performance and our results provide practical guidelines for these choices. We propose a new adaptive and hybrid procedure which is stable and faster while maintaining the same theoretical properties.

7.1 Introduction

We often face incomplete data such as censored and/or truncated data in survival analysis, and unobservable latent variables in mixed effects and errors-in-variables models. It is difficult to compute estimates of parameters such as maximum likelihood estimates (MLE) with incomplete data because the likelihood usually involves intractable integrals. For example, the likelihood function of the generalized linear mixed models (GLMM) involves an integral with respect to the random effects which usually does not have a closed form for non-Gaussian observations. Laplace approximation may lead to large bias and inconsistent estimates [3, 7, 11]. McCulloch [13] modified the EM and the Newton–Raphson algorithms for fitting GLMMs with integrals approximated by Monte Carlo (MC) methods. A drawback of the MC approach is that the iterative scheme converges to a random variable rather than to the expected fixed point. As a consequence, it is difficult

Y. Jiang (✉)
Capital One Financial Corp., 15000 Capital One Dr., Richmond, VA 23238, USA
e-mail: Yihua.Jiang@capitalone.com

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_7,

to decide the MC sample size and a stopping rule for such an iterative procedure. Booth and Hobert [2] pointed out these potential problems and proposed an empirical approach to increase the MC sample size. However, it is unclear if their algorithm converges in theory.

MCMCSAA proposed by [5] (Gu and Kong, 2000, Personal communication) for incomplete data is a major breakthrough which guarantees convergence to the expected fixed points in theory. MCMCSAA uses stochastic approximation (SA) to update the parameters and Markov chain Monte Carlo (MCMC) to approximate integrals at each iteration. In this article we investigate the performance of MCMCSAA using simulations. Our goal is to reveal some practical issues involved in the implementation of MCMCSAA. We find that MCMCSAA can be sensitive to the choices of (a) the initial values, (b) the MCMC sample size and step-size of the parameter update, and (c) a matrix whose expected value approximates the Hessian of the target function. We provide recommendations for implementation and propose a hybrid algorithm which is stable and considerably faster.

We review the MCMCSAA in Sect. 7.2. We conduct simulations to illustrate potential pitfalls in the implementation of the MCMCSAA in Sect. 7.3. We propose a hybrid algorithm in Sect. 7.4. Concluding remarks are made in Sect. 7.5.

7.2 Markov Chain Monte Carlo Stochastic Approximation Algorithms

In this section we briefly review MCMCSAA proposed by [5] (Gu and Kong, 2000, Personal communication). See also [6, 9, 10] (Gu and Zhu, 2002, Personal communication).

Suppose that we want to solve the following estimating equations

$$\vec{h}(\vec{\theta}) = \mathbf{0}, \tag{7.1}$$

where $\vec{\theta}$ is a vector of parameters and $\vec{h}(\vec{\theta})$ is a vector valued function that can be written as the expectation of a function $\vec{H}(\vec{\theta}, \vec{e})$ with respect to a random vector $\vec{e}$ with density function $f_{\vec{e}}(\vec{e})$:

$$\vec{h}(\vec{\theta}) = E_{\vec{e}}[\vec{H}(\vec{\theta}, \vec{e})] = \int \vec{H}(\vec{\theta}, \vec{e}) f_{\vec{e}}(\vec{e}) d\vec{e}. \tag{7.2}$$

In survival analysis, for example, data might be right censored. The score function can be written in form (7.2) [14]. Another example is a GLMM when the random effects are considered as missing data. Integrating respect to the vector of random effects leads to form (7.2) [5, 13, 15].

To apply the MCMCSAA, one needs to find a matrix $I(\vec{\theta}, \vec{e})$ such that $E_{\vec{e}}[I(\vec{\theta}, \vec{e})] \approx \partial\vec{h}/\partial\vec{\theta}$ in the neighborhood of the solution. We consider the following three different choices of $I(\vec{\theta}, \vec{e})$:

$$
\begin{aligned}
I_1(\vec{\theta},\vec{e}) &= -\partial\vec{H}(\vec{\theta},\vec{e})/\partial\vec{\theta},\\
I_2(\vec{\theta},\vec{e}) &= -\partial\vec{H}(\vec{\theta},\vec{e})/\partial\vec{\theta} - \vec{H}(\vec{\theta},\vec{e})\vec{H}(\vec{\theta},\vec{e})^T,\\
I_3(\vec{\theta},\vec{e}) &= -\partial\vec{H}(\vec{\theta},\vec{e})/\partial\vec{\theta} - \vec{H}(\vec{\theta},\vec{e})\vec{H}(\vec{\theta},\vec{e})^T + \mathrm{E}_{\vec{e}}[\vec{H}(\vec{\theta},\vec{e})]\mathrm{E}_{\vec{e}}[\vec{H}(\vec{\theta},\vec{e})]^T.
\end{aligned}
\tag{7.3}
$$

Whenever the meaning is clear we will drop the dependence on $\vec{\theta}$ and $\vec{e}$ from all I-matrices. Also for brevity we will often refer to I_1, I_2, or I_3 when we actually mean the algorithm based on I_1, I_2, or I_3. I_1 used by [1], is usually well-behaved. For example, it is positive definite for convex target functions. Gu and Kong (2000, Personal communication) proposed I_2 and claimed that it is more efficient. I_3 motivated by the actual derivative of $\vec{h}(\vec{\theta})$ with respect to $\vec{\theta}^T$ [12] is new. Convergence of the algorithms are guaranteed for all three choices of I-matrices [1, 5]. (Gu and Kong, 2000, Personal communication).

Let $\{\gamma_k, k \geq 1\}$ be a sequence of real numbers, and $\{m_k, k \geq 1\}$ be a sequence of positive integers which satisfy the following conditions:

(C1) $\gamma_k \leq 1$ for all k

(C2) $\sum_{k=1}^{\infty} \gamma_k = \infty$

(C3) $\sum_{k=1}^{\infty} \gamma_k^{1+\epsilon}/m_k < \infty$, for some $\epsilon \in (0, 1)$

(C4) $\sum_{k=1}^{\infty} |\gamma_k/m_k - \gamma_{k-1}/m_{k-1}| < \infty$

At iteration k, a MCMC sample of size m_k with equilibrium distribution $f_{\vec{e}}(\vec{e})$ is drawn. See [4] for an overview of the Metropolis–Hastings algorithm. Let $\vec{e}_k^{(1)}, \ldots, \vec{e}_k^{(m_k)}$ be the MCMC sample of size m_k after some burn-in. The MCMCSAA updates the parameter vector $\vec{\theta}$ and a matrix Γ as follows:

$$
\begin{aligned}
\Gamma_k &= (1-\gamma_k)\Gamma_{k-1} + \gamma_k \bar{I}_k,\\
\vec{\theta}_k &= \vec{\theta}_{k-1} + \Gamma_k^{-1}\bar{\vec{H}}_k,
\end{aligned}
\tag{7.4}
$$

where

$$
\bar{\vec{H}}_k = \sum_{j=1}^{m_k} \vec{H}(\vec{\theta}_{k-1}, \vec{e}_k^{(j)})/m_k, \tag{7.5}
$$

$$
\bar{I}_k = \sum_{j=1}^{m_k} I(\vec{\theta}_{k-1}, \vec{e}_k^{(j)})/m_k. \tag{7.6}
$$

The I matrix in (7.6) may take any form in (7.3). Γ_k acts as a proxy of the Hessian matrix and is updated as parameters. γ_k is the step-size of the parameter updates. The innovation of MCMCSAA is the introduction of two sequences $\{\gamma_k\}$ and $\{m_k\}$. For a constant m_k and $\gamma_k = 1$, the estimate of $\vec{\theta}$ will maintain a certain amount

of variation along iterations due to estimating $\vec{H}$ and I by averaging over random samples. By increasing the sample size m_k, decreasing step-size γ_k, or both, the MCMCSAA decreases the variation in $\vec{\theta}$ as iteration increases. Gu and Kong (2000, Personal communication) have shown that under some regularity conditions and when m_k and γ_k satisfy conditions (C1)–(C4), $\vec{\theta}_k$ converges to the solution of (7.1) almost surely. The following are three combinations of m_k and γ_k that satisfy conditions (C1)–(C4):

G1 $\gamma_k = 1$ and $m_k = m_0 + k^2$

G2 $\gamma_k = 1/k$ and $m_k = m_0$

G3 $\gamma_k = 1/\sqrt{k}$ and $m_k = m_0 + k$

where m_0 is the starting MCMC sample size. We see that when $\gamma_k = 1$, m_k needs to increase quadratically. When m_k is a constant, γ_k needs to decrease in the order of $1/k$. G3 is a compromise between G1 and G2.

The result of (Gu and Kong, 2000, Personal communication)provides the theoretical justification for the MCMCSAA. However, few research has been done to investigate its performance in practice. The simulation results in the next section are intended to show some potential problems that may occur in practice, and to compare different choices of parameters of the algorithm. We propose a new hybrid procedure and conduct more simulations in Sect. 7.4.

7.3 Simulations

To implement the MCMCSAA, one needs to decide (a) the initial values $\vec{\theta}_0$, (b) the MCMC sample size m_k and step-size of the parameter update γ_k, and (c) the form of the I-matrix. We will use simulations to show that the stability of the algorithm and the speed of convergence depends critically on all three choices.

We use the following simple model to illustrate challenges and potential pitfalls involved in the implementation of the MCMCSAA. Same issues arise in more complex models. Consider an experiment with $n = 10$ binary observations from each of $m = 20$ subjects. Binary observations are generate from the following simple GLMM:

$$\begin{aligned} &P(y_{ij} = 1|b_i) = \exp(b_i)/(1 + \exp(b_i)), \\ &b_i \overset{iid}{\sim} \mathrm{N}(0, \theta), \;\; i = 1, \ldots, m; \;\; j = 1, \ldots, n, \end{aligned} \tag{7.7}$$

where b_i are random effects. Note that there is no fixed effects in model (7.7). The problems we are going to show in this section are usually associated with the estimation of variance components such as θ, while the estimation of fixed effects is stable and converges quickly. Thus we set them to zero to simplify the exposition.

Let $\vec{y} = (y_{11}, \cdots, y_{1n}, \cdots, y_{m1}, \cdots, y_{mn})^T$ and $\vec{b} = (b_1, \cdots, b_m)$. The log-marginal distribution of $\vec{y}$ is

$$l_{\vec{y}}(\vec{y}; \theta) = \log \int f_{\vec{y}|\vec{b}}(\vec{b}; \theta) \exp\{l_{\vec{b}}(\vec{b}; \theta)\} d\vec{b}, \tag{7.8}$$

where $f_{\vec{y}|\vec{b}}(\vec{b}; \theta)$ is the conditional distribution of $\vec{y}$ conditional on $\vec{b}$ and $l_{\vec{b}}(\vec{b}; \theta)$ is the log-density of $\vec{b}$. It can be shown that under some regularity conditions,

$$\frac{\partial l_{\vec{y}}(\vec{y}, \theta)}{\partial \theta} = \int \frac{\partial l_{\vec{b}}(\vec{b}; \theta)}{\partial \theta} f_{\vec{b}|\vec{y}}(\vec{b}) d\vec{b} = \mathrm{E}_{\vec{b}|\vec{y}} \frac{\partial l_{\vec{b}}(\vec{b}; \theta)}{\partial \theta},$$

where $f_{\vec{b}|\vec{y}}$ is the posterior density of $\vec{b}$ given $\vec{y}$. Thus we can apply MCMCSAA with $\vec{e} = \vec{b}|\vec{y}$ and $\vec{H}(\theta, \vec{b}) = \partial l_{\vec{b}}(\vec{b}, \theta)/\partial\theta$. It is easy to check that for model (7.7),

$$\begin{aligned}
\vec{H}(\theta, \vec{b}) &= -m/(2\theta) + \sum_{i=1}^{m} b_i^2/(2\theta^2), \\
I_1(\theta, \vec{b}) &= -m/(2\theta^2) + \sum_{i=1}^{m} b_i^2/\theta^3, \\
I_2(\theta, \vec{b}) &= I_1(\theta, \vec{b}) - \left[-m/(2\theta) + \sum_{i=1}^{m} b_i^2/(2\theta^2) \right]^2, \\
I_3(\theta, \vec{b}) &= I_2(\theta, \vec{b}) + (\mathrm{E}_{\vec{b}|\vec{y}} \vec{H}(\theta, \vec{b}))^2.
\end{aligned} \tag{7.9}$$

Since θ is non-negative, we set $\theta_k = \theta_{k-1}$ if the update $\theta_{k-1} + \gamma_k \boldsymbol{\Gamma}_k^{-1} \bar{\mathbf{H}}_k < 0$.

We estimate $\mathrm{E}_{\vec{b}|\vec{y}} \vec{H}(\theta, \vec{b})$ by $\bar{H}(\theta, \vec{b}) = \sum_{j=1}^{m} \vec{H}(\theta, \vec{b}^{(j)})$, where $\vec{b}^{(j)}$ are MCMC samples. We use the single component Metropolis–Hastings algorithm to sample from $f_{\vec{b}|\vec{y}}(\vec{b})$ with the length of burn-in set to 300 [4, 8]. Specifically, at the pth iteration of the MCMC, we update each element $b_1^{(p)}, \ldots, b_m^{(p)}$ one by one as follows: generate X from $\mathrm{N}(b_i^{(p-1)}, 0.5\theta_{k-1})$ as a candidate for $b_i^{(p)}$ and U from Unif(0, 1); set $b_i^{(p)} = X$ if $U \le \min\{1, \pi(X)/\pi(b_i^{(p-1)})\}$ and $b_i^{(p)} = b_i^{(p-1)}$ otherwise.

We simulate data with three true values of θ: 0.5, 1.0 and 2.0. Denote θ_0 as the initial value and θ_{MLE} as the MLE. We compute θ_{MLE}'s numerically by maximizing (7.8) using integration and optimization functions in R (www.r-project.org). We consider the following combinations of parameters for the algorithm:

- Initial values θ_0: $0.5\theta_{MLE}$, θ_{MLE} and $2\theta_{MLE}$
- I-matrices: I_1, I_2 and I_3
- γ_k and m_k: G1, G2 and G3
- m_0: 30 and 300

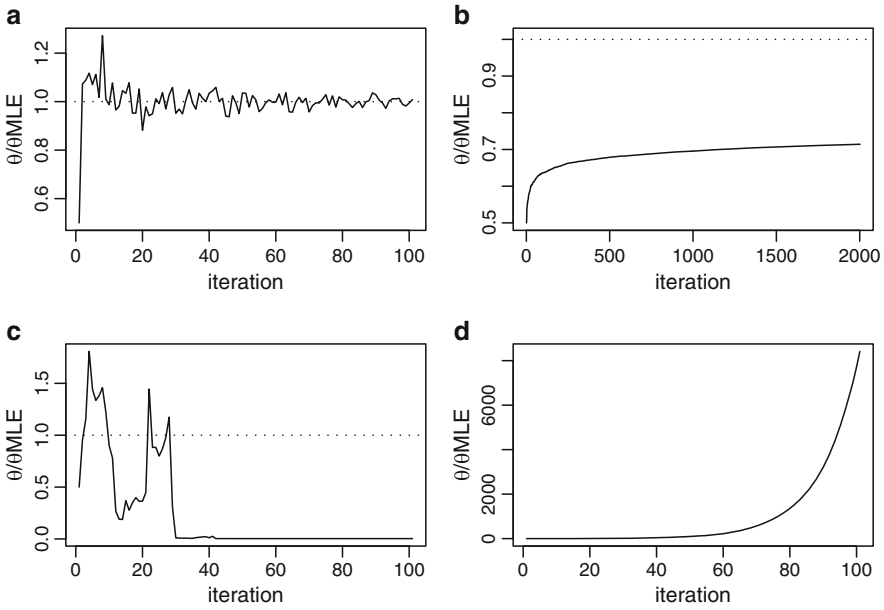


Fig. 7.1 Four patterns of sequences divided by θ_{MLE}

We have also tried to set both m_0 and burn-in period to 10,000, the convergence behavior remains the same. For each combination of parameters, we repeat simulation 100 times. Following (Gu and Kong, 2000, Personal communication), we stop the algorithm after 50 iterations for G1, 1,000 iterations for G2, and 250 iterations for G3.

We first look at the convergence behavior. All iteration sequences follow one of the following four patterns: (a) converged to the MLE, (d) approaching but not yet converged to the MLE, (c) converges to zero, and (d) diverges to infinity. We show a typical example for each pattern in Fig. 7.1.

The curve of sequences usually becomes smoother and smoother along the iterations. However, depending on the choice of γ_k and m_k, random variation persists even after a large number of iterations. The following definitions allow the estimates to have small variations around the MLE. Let $\bar{\theta}_5$ be average of the estimate at the last 5 iterations. For any sequence, we define the following three states:

(S1) *converged*, if $\left|(\bar{\theta}_5 - \theta_{MLE})/(\theta_{MLE} + 1)\right| < 0.05$
(S2) *diverged*, if $\left|(\bar{\theta}_5 - \theta_{MLE})/(\theta_{MLE} + 1)\right| > 1$, or $\left|(\bar{\theta}_5 - \theta_{MLE})/(\theta_{MLE} + 1)\right| \geq 0.05$ and $\left|\bar{\theta}_5/\theta_{MLE}\right| < 0.05$
(S3) *not converged*, otherwise

Not converged sequences (S3) consists of two classes: sequences that are converging but are not close enough to the MLE and sequences that are diverging but are not far enough from the MLE. Table 7.1 shows the convergence behavior. An entry i/j

Table 7.1 Number of simulation replications that are converged or diverged

θ	θ_0		$m_0 = 30$			$m_0 = 300$		
			G1	G2	G3	G1	G2	G3
		$\mathbf{I}_1$	78/0	0/0	72/0	78/0	0/0	75/0
	$0.5\theta_{MLE}$	$\mathbf{I}_2$	8/85	55/26	43/51	11/86	74/17	61/32
		$\mathbf{I}_3$	38/47	53/6	70/15	50/38	80/6	75/15
		$\mathbf{I}_1$	79/0	64/0	85/0	83/0	92/0	89/0
0.5	$1\theta_{MLE}$	$\mathbf{I}_2$	5/93	55/36	32/64	21/75	84/12	62/30
		$\mathbf{I}_3$	38/54	65/8	60/18	53/36	85/9	71/14
		$\mathbf{I}_1$	82/0	46/0	87/0	87/0	55/0	90/0
	$1.5\theta_{MLE}$	$\mathbf{I}_2$	4/95	27/65	16/80	11/88	46/47	24/71
		$\mathbf{I}_3$	34/54	63/9	50/36	50/44	83/5	71/22
		$\mathbf{I}_1$	95/0	0/0	94/0	97/0	0/0	94/0
	$0.5\theta_{MLE}$	$\mathbf{I}_2$	10/85	65/22	49/44	22/77	77/17	72/24
		$\mathbf{I}_3$	79/19	75/3	83/6	79/17	97/1	92/2
		$\mathbf{I}_1$	97/0	86/0	98/0	95/0	100/0	99/0
1	$1\theta_{MLE}$	$\mathbf{I}_2$	9/87	66/29	58/38	51/48	96/3	89/9
		$\mathbf{I}_3$	64/33	81/6	81/7	81/18	98/0	94/6
		$\mathbf{I}_1$	96/0	77/0	98/0	97/0	89/0	98/0
	$1.5\theta_{MLE}$	$\mathbf{I}_2$	3/95	28/64	26/72	6/93	38/54	30/69
		$\mathbf{I}_3$	67/31	82/6	86/9	80/19	89/3	91/7
		$\mathbf{I}_1$	98/0	0/0	98/0	98/0	0/0	99/0
	$0.5\theta_{MLE}$	$\mathbf{I}_2$	11/88	66/24	58/40	32/66	53/42	49/49
		$\mathbf{I}_3$	84/14	95/1	97/1	91/9	100/0	100/0
		$\mathbf{I}_1$	98/0	97/0	99/0	99/0	100/0	100/0
2	$1\theta_{MLE}$	$\mathbf{I}_2$	12/87	85/13	58/40	58/42	99/1	95/4
		$\mathbf{I}_3$	85/15	96/0	93/5	91/8	100/0	99/1
		$\mathbf{I}_1$	98/0	87/0	99/0	100/0	98/0	100/0
	$1.5\theta_{MLE}$	$\mathbf{I}_2$	8/91	35/64	19/81	5/95	16/84	12/88
		$\mathbf{I}_3$	82/17	78/9	91/8	92/5	82/3	98/0

indicates that there are i converged sequences, j diverged sequences, and $100-i-j$ not converged sequences.

It is clear that the convergence behavior depends on θ_0 as well as the choices of I, γ_k and m_k. As always, a good initial value is important. I_1 is stable as expected: most all sequences based on I_1 converged and none of them diverged. A closer look indicates that all not converged sequences are actually converging slowly. In contrast to I_1, I_2 and I_3 are unstable with many sequences diverged, regardless of the choice of θ_0, I, γ_k and m_k. It is important to note that the problem persists even when the initial value is set to the target value θ_{MLE}. I_2 has more diverged sequences than I_3. Convergence also depends on the specific combination of I_1, I_2, I_3 and G1, G2, G3. For example, I_2 always performs best with G2. And when θ_0 is not close to θ_{MLE}, I_1 performs better with G1 than with G2.

To illustrate the cause of divergence when I_2 and I_3 are used, for a typical diverged case, we plot θ and Γ for the first 10 iterations in Fig. 7.2. We see that

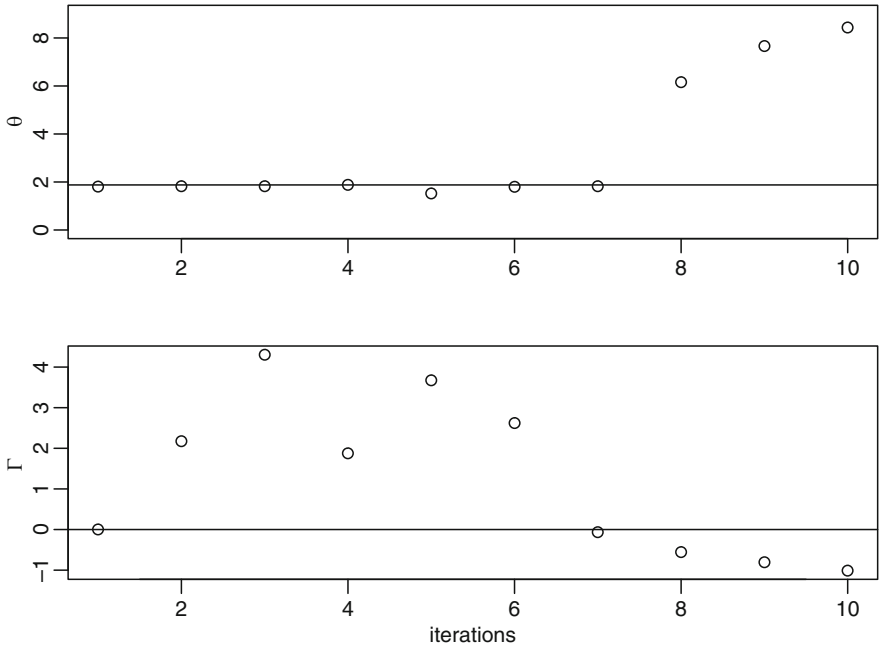


Fig. 7.2 Effect of negative Γ on estimate of θ. The *bottom plot* shows the value of Γ based on I_2. The *top plot* shows the estimate of θ at the corresponding iteration, and the *solid line* represents the MLE

from iteration 7, due to MCMC simulation error, the Γ values are negative (i.e., the Γ matrices are not positive definite). This turns θ away from the MLE. In most cases sequences with such problems do not recover and diverge to infinity or converge to zero. Zhu and Lee [15] noticed the same problem and proposed a hybrid algorithm.

Next, we look at the rate of convergence. In Table 7.1 we see that for all combinations, the algorithm based on G2 has the most sequences that are classified as *not converged*. To compare G1 and G2, we show two typical sequences of estimates based on I_1 with $\theta_0 = \theta_{MLE}$ and $\theta_0 = 0.5\theta_{MLE}$ in Fig. 7.3.

The left plot shows that θ/θ_{MLE} are close to 1 for both G1 ad G2, which means that both sequences settle around the MLE, but the curve based on G2 is much smoother than the curve based on G1. The right plot shows that the sequences based on G1 converged to the MLE in fewer than 50 iterations, whereas the sequence based on G2 has not reached the level of the MLE after 1,000 iterations.

Plots under other settings are similar. We conclude that if the starting value is far from the MLE, G1 reaches the level of the MLE faster than G2. On the other hand if the estimate is already at the level of the MLE, G2 settles down much faster than G1. These results are not surprising since for G2 the step-size of the parameter update γ_k is $1/k$. Thus after only a few iteration γ_k is very small and does not allow a large change in θ anymore. It takes many iterations to reach the level of the MLE.

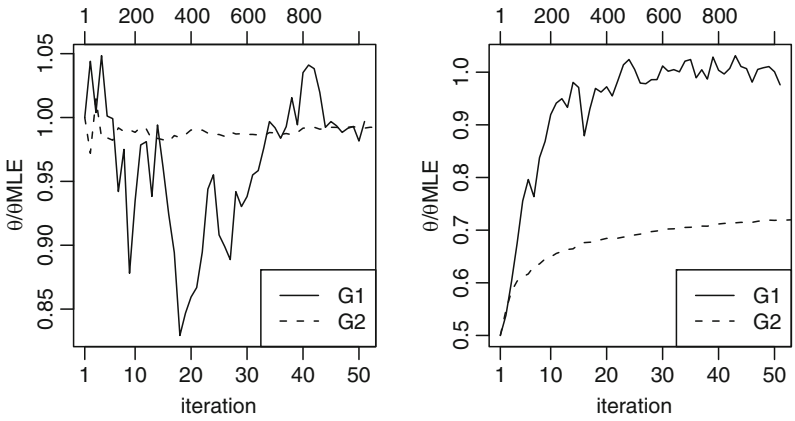


Fig. 7.3 Ratios between the estimates and θ_{MLE} for sequences based on G1 and G2. The iteration numbers of G1 are marked at the *bottom* of the x-axis and iteration numbers of G2 are marked at the *top*

For G1, γ_k is always 1. Therefore, it can reach the level of the MLE quickly, but remain to have random variations after the level of the MLE has been reached. G3 is a compromise between G1 and G2 (not shown).

Above simulation results illustrate advantages and disadvantages of different combinations of the I matrices, step size γ_k and MCMC sample size m_k. These results can be used as a guide to make an algorithm stable and efficient. Overall, we recommend I_1 for stability. One may use G1 and then switch to G2 after the sequence has reached the target level (Sect. 7.4). When I_2 or I_3 are used, one needs to check the properties of the estimates, for example, if the Γ matrices are positive definite. One can start with I_1 and switch to I_2 or I_3 after a number of iterations [15]. This would combine the stability of I_1 in the beginning of the algorithm with the faster convergence of I_2 or I_3 at the end of the algorithm.

7.4 A Hybrid Algorithm

We consider I_1 only in this section. Based on simulations in the previous section, we found that G1 reaches the MLE quickly and G2 reduces variation in the estimates quickly after the sequence reached the level of the MLE. Therefore, we consider the following class of algorithms:

$$\gamma_k = k^{-t_k} \text{ and } m_k = m_0 + k^{2(1-t_k)}, \quad 0 \le t_k \le 1. \tag{7.10}$$

Special cases include G1 with $t_k = 0$, G3 with $t_k = 0.5$, and G2 with $t_k = 1$. Note that the sequences need to satisfy conditions (C1)–(C4).

We would like t_k to be close to zero for small k and close to one for large k. The idea of our procedures is to select γ_k and m_k adaptively according to the actual convergence behavior. There could be many adaptive schemes. We consider a simple approach by fitting a regression line through the last K estimates of $\boldsymbol{\theta}$. If the slope is zero, it is likely that the estimates have reached the level of the MLE and the variation of $\boldsymbol{\theta}$ around the regression line is likely due to MCMC sampling. In this case we want to reduce the variation using a combination of γ_k and m_k similar to G2. If the slope is not zero, i.e., the sequence is still increasing or decreasing, it is desirable to choose γ_k and m_k such that the estimates reach the level of MLE quickly. One such choice is G1. Let r_k be the correlation coefficient between the estimate of $\boldsymbol{\theta}$ at the previous K iterations and their corresponding iteration numbers, and $\hat{r}_k$ be the sample correlation coefficient. Zero r_k corresponds to the zero slope of the regression line. We use t-test to test hypothesis $H_0 : r_k = 0$ *v.s.* $H_1 : r_k \neq 0$. Specifically, H_0 is rejected when $\left|\hat{r}_k / \sqrt{(1-\hat{r}_k^2)/(K-2)}\right| \geq t_{\alpha/2,K-2}$ where α is the significant level and $t_{\alpha/2,K-2}$ is the $1-\alpha/2$ percentile of the t-distribution with $K-2$ degrees of freedom. This suggests the following algorithms:

G4 For $k > K$, $t_k = 1 - r_k^2$

G5 For $k > K$, $t_k = (1-r_k^2) I(|\hat{r}_k / \sqrt{(1-\hat{r}_k^2)/(K-2)}| < t_{\alpha/2,K-2})$

G6 For $k > K$, $t_k = I(|\hat{r}_k / \sqrt{(1-\hat{r}_k^2)/(K-2)}| < t_{\alpha/2,K-2})$

G1 is used for the first K iterations in G4, G5 and G6. To implement algorithms G4, G5 and G6, we need to decide the number K, that is, the last K estimates to be used to compute the correlation coefficients.

For the same simulated data we apply algorithms G4, G5 and G6 with four different choices of K: 10, 20, 30 and 40. We stop the hybrid algorithms after the 50th iteration which is consistent with the number used for G1 in the previous section. Table 7.2 presents the number of converged sequences out of 100 simulation replications. There is no diverged sequences.

Table 7.2 indicates that the convergence behavior is insensitive to the choice of K so long it is large enough ($\geq$10). To compare the convergence rate with G1, we plot a typical replication of hybrid algorithms with $K = 20$ in Fig. 7.4. We see that the curves based on hybrid algorithms are much smoother than those based on G1. However, it is unclear whether 50 iterations are enough for the sequences to reach the θ_{MLE}. We need to consider stopping rules.

Different convergence criteria have been used in practice. One simple rule is to stop the algorithm when $|\theta_{k+1} - \theta_k| < \delta$ where δ is a preassigned number. Gu and Zhu [6] proposed a two-stage algorithm where the first stage stops at K_1 where $K_1 = \inf\left\{K \geq K_0 : || \sum_{k=K-K_0+1}^{K} \frac{sgn(\theta_k - \theta_{k-1})}{K_0} || \leq \delta_1\right\}$, K_0 and δ_1 are preassigned constants and sgn is the signum function of a real number x which returns 1 when $x > 0$, 0 when $x = 0$ and -1 when $x < 0$. The second stage utilizes the trace of Γ matrix and H matrix to decide the stopping rule. We use the variance along iterations to measure whether the sequence is stable or not. Specifically we consider the following two stopping rules:

Table 7.2 Number of simulation replications that are converged

θ_0		True value θ								
		0.50			1.00			2.00		
		G4	G5	G6	G4	G5	G6	G4	G5	G6
$0.5\theta_{MLE}$	$K = 10$	64	76	73	90	94	96	96	98	99
	$K = 20$	65	78	75	89	94	97	97	97	98
	$K = 30$	72	78	76	89	94	95	97	98	100
	$K = 40$	69	75	77	93	95	96	98	99	97
$1\theta_{MLE}$	$K = 10$	83	83	84	90	95	95	99	98	99
	$K = 20$	73	80	83	93	96	94	97	100	98
	$K = 30$	73	82	81	91	93	96	100	99	100
	$K = 40$	82	85	79	96	94	96	99	99	99
$1.5\theta_{MLE}$	$K = 10$	86	84	79	92	93	99	97	99	99
	$K = 20$	75	86	87	91	96	98	97	98	99
	$K = 30$	75	87	86	96	93	99	99	99	100
	$K = 40$	84	83	84	93	94	95	100	99	98

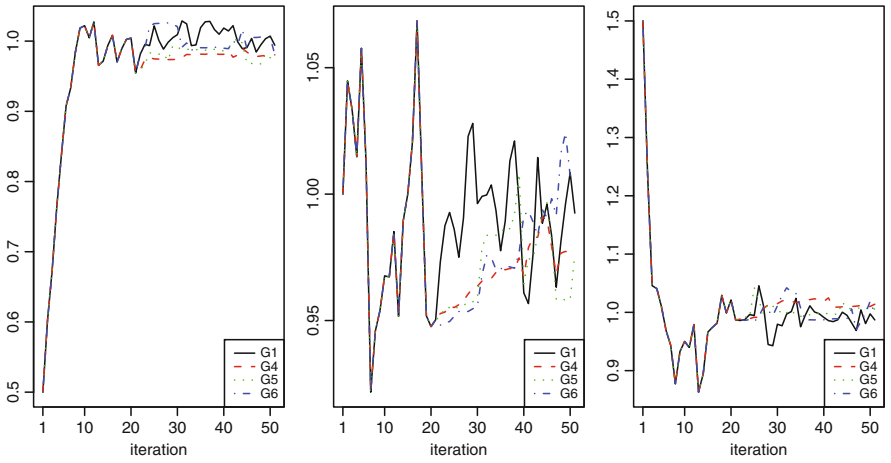


Fig. 7.4 Ratios between the estimates and θ_{MLE} for sequences based on G1, G4, G5 and G6

I At the kth iteration, stop if $\frac{|\theta_k - \theta_{k-1}|}{\sqrt{v}+\delta_1} < \delta_2$ where δ_1, δ_2 are predetermined numbers and v is the sample variance of $\{\theta_0, ...\theta_k\}$.

II Since Γ_k^{-1} is an estimate of the covariance matrix of $\hat{\theta}$, therefore, instead of v, we stop if $\frac{|\theta_k - \theta_{k-1}|}{\sqrt{\Gamma_k^{-1}}+\delta_1} < \delta_2$.

For the same simulation settings, we apply algorithms G1–G6 using above two stopping rules. For both rules, we set $\delta_1 = 0.001$ and $\delta_2 = 0.001$. We compute k as the average of number of iterations needed to satisfy the stopping criterion, *conv*. as converged cases after 600 iterations, *diff* as the average of the difference between

Table 7.3 Convergence behavior of G1, G3, G4, G5, and G6 under stopping rule I, *diff* are multiplied by 1,000 for ease of presentation

θ	θ_0		G1	G4	G5	G6	G3
0.5	$0.5\theta_{MLE}$	k	118.59	50.12	55.02	53.53	77.40
		conv.	87	68	78	78	47
		diff	6.672	19.60	12.55	14.99	38.5
		CPU (s)	66.12	1.34	4.38	4.67	2.64
0.5	$1\theta_{MLE}$	k	216.40	80.38	82.44	71.86	179.42
		conv.	92	82	83	84	84
		diff	2.677	10.66	8.424	9.49	7.407
		CPU (s)	390.82	2.71	16.49	9.80	6.77
0.5	$1.5\theta_{MLE}$	k	131.41	45.04	53.75	51.84	87.12
		conv.	92	73	78	84	70
		diff	5.881	13.90	11.46	10.16	15.23
		CPU (s)	113.02	0.94	4.09	4.10	3.04
1	$0.5\theta_{MLE}$	k	125.4	45.8	49.49	43.13	66
		conv.	95	87	95	91	65
		diff	7.588	22.13	14.15	18.86	43.98
		CPU (s)	84.55	1.03	2.70	2.03	2.25
1	$1\theta_{MLE}$	k	289.66	90.9	89.66	77.59	222.02
		conv.	98	94	97	98	99
		diff	3.188	12.41	9.512	11.95	10.35
		CPU (s)	828.94	3.03	17.29	10.797	8.9
1	$1.5\theta_{MLE}$	k	156.46	90.9	56.05	49.42	88.16
		conv.	98	94	95	93	89
		diff	6.854	20.41	17.67	17.10	20.92
		CPU (s)	199.19	1.11	4.56	3.38	3.11
2	$0.5\theta_{MLE}$	k	132.87	44.82	49.33	43.01	68.16
		conv.	100	97	96	98	84
		diff	13.46	43.17	34.76	31.84	57.68
		CPU (s)	111.41	0.93	2.92	1.88	2.31
2	$1\theta_{MLE}$	k	272.06	96.15	90.54	79.69	239.70
		conv.	99	97	100	100	94
		diff	8.389	26.86	17.34	17.89	20.89
		CPU (s)	720.95	3.27	14.30	11.57	9.47
2	$1.5\theta_{MLE}$	k	162.57	49.47	55.96	47.83	80.69
		conv.	100	94	97	99	94
		diff	15.68	40.31	30.07	30.78	37.05
		CPU (s)	195.29	1.07	4.44	2.99	2.80

estimates and the θ_{MLE} among converged sequences, and *CPU* as the average CPU time in seconds for one simulation replication.

Tables 7.3 and 7.4 show convergence behavior under rules I and II respectively. We did not include the convergence behavior of G2 because G2 performs poorly under these two stopping rules. G2 makes the algorithm stop early even though the iteration is far away from the targeted θ_{MLE}.

Table 7.4 Convergence behavior of G1, G3, G4, G5 and G6 under stopping rule II, *diff* are multiplied by 1,000 for ease of presentation

θ	θ_0		G1	G4	G5	G6	G3
0.5	$0.5\theta_{MLE}$	k	72.58	38.59	43.95	39.79	43.32
		conv.	83	59	70	70	24
		diff	9.562	23.15	17.69	18.51	59.47
		CPU (s)	14.02	1.55	3.81	2.55	1.40
0.5	$1\theta_{MLE}$	k	78.08	32.16	34.53	32.79	35.41
		conv.	88	72	75	77	73
		diff	7.743	16.12	15.23	13.35	14.72
		CPU (s)	16.28	1.20	1.76	1.68	1.16
0.5	$1.5\theta_{MLE}$	k	68.98	32	35.08	34.32	38.89
		conv.	88	74	77	76	57
		diff	8.929	16.53	13.36	14.36	26.09
		CPU (s)	11.97	1.18	1.85	1.77	1.27
1	$0.5\theta_{MLE}$	k	71.37	33.06	38.46	36.41	38.48
		conv.	96	81	91	91	41
		diff	11.90	28.44	21.90	24.23	66.73
		CPU (s)	13.83	1.26	2.17	1.96	1.24
1	$1\theta_{MLE}$	k	63.86	30.2	33.91	32.68	35.98
		conv.	94	87	91	89	92
		diff	11.16	23.01	18.18	20.98	19.49
		CPU (s)	10.07	1.12	1.68	1.61	1.17
1	$1.5\theta_{MLE}$	k	65.60	30.88	31.86	33.01	36.02
		conv.	98	92	91	88	84
		diff	10.75	21.11	21.80	24.16	27.19
		CPU (s)	12.44	1.14	1.48	1.64	1.17
2	$0.5\theta_{MLE}$	k	67.01	34.21	34.19	32.87	35.21
		conv.	98	94	96	97	61
		diff	22.34	48.81	43.22	42.02	13.57
		CPU (s)	11.25	1.28	1.66	1.62	1.14
2	$1\theta_{MLE}$	k	65.75	29.83	29.85	27.5	32.81
		conv.	100	96	96	97	95
		diff	21.46	47.51	39.91	48.30	37.52
		CPU (s)	10.89	1.10	1.30	1.13	1.07
2	$1.5\theta_{MLE}$	k	68.42	29.67	32.35	30.73	33.84
		conv.	100	95	99	96	99
		diff	23.19	50.40	41.64	47.49	51.93
		CPU (s)	12.85	1.10	1.53	1.38	1.10

First, we see that stopping rule I is much more time-consuming than stopping rule II for G1, but not for the hybrid algorithms. And when $\theta_0 = \theta_{MLE}$, G1 with stopping rule I is very slow to converge, the average CPU time is almost 200 times of those for hybrid algorithms.

Overall, G1 has the highest *converged* cases among all the algorithms. On the other hand, G1 is always the most time-consuming one. G3 usually has the lowest *converged* cases especially when θ_0 is far away from the θ_{MLE}. G3 is the tradeoff

between G1 and G2 and the sequences based on G3 is much smoother than G1. So if we use the variation of sequences as the stopping rule, G3 may stop prematurely. The hybrid algorithms overcome this problem by switching between G1 and G2, so they give better *converged* rate than G3 even though the average CPU time for hybrid algorithms and G3 are at the same level.

Now we look at the average difference between estimates and θ_{MLE}. Under stopping rule I, when $\theta = 0.5$ and $\theta = 1$, the average difference is at 10^{-3} level for G1, and at 10^{-2} level for hybrid algorithms; when $\theta = 2$, the average difference is at 10^{-2} level for both G1 and hybrid algorithms. Under stopping rule II, when $\theta = 0.5$, the average difference is at 10^{-3} level for G1, and at 10^{-2} level for hybrid algorithms; when $\theta = 1$ and $\theta = 2$, the average difference is at 10^{-2} level for both G1 and hybrid algorithms. The average difference based on G1 are always smaller then the hybrid algorithms. However, the CPU time for G1 is between 10 to 200 times of that for the hybrid algorithms.

7.5 Conclusions

MCMCSAA is a powerful tool to compute estimates of parameters with incomplete data. It has attractive theoretical properties. However, due to its complexity and flexibility, the implementation of MCMCSAA is far from straightforward and caution needs to be exercised.

The MCMCSAA involves several parameters including the form of the I-matrix, the MCMC sample size m_k and step-size of the parameter update γ_k. We have shown that the performance of MCMCSAA depends critically on all three parameters. We recommend I_1 over I_2 and I_3 for stability. When I_2 or I_3 is used, one should check the property of the H matrix to avoid divergence. There can be many choices for the combination of the MCMC sample size m_k and step-size of the parameter update γ_k. Different choices represent different compromise between convergence rate and CPU time. There is no ultimate best algorithm for both precision and efficiency. Therefore, different choice should be chosen for different purpose. In general, the proposed hybrid algorithms are stable and fast.

References

1. Benveniste A, Métivier M, Priouret B (1987) Adaptive algorithms and stochastic approximations: Theory and applications of signal detection and processing and pattern recognition (French). Masson, Paris
2. Booth J, Hobert J (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo em algorithm. J R Stat Soc Series B 61:265–285
3. Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88:9–25
4. Gilks WR, Richardson S, Spiegelhalter DJE (1996) Markov chain Monte Carlo in practice. Chapman and Hall, London

5. Gu MG, Kong FH (1998) A stochastic approximation algorithm with Markov chain Monte–Carlo for imcomplete data estimation problems. Proc Natl Acad Sci 95:7270–7274
6. Gu MG, Zhu HT (2001) Maximum likelihood estimation for spatial models by Markov chain Monte–Carlo stochastic approximation. J R Stat Soc Series B 63:339–355
7. Jiang J (1998) Consistent estimators in generalized linear mixed models. J Am Stat Assoc 93:720–729
8. Karcher P, Wang Y (2002) Generalized nonparametric mixed effects models. J Comput Graph Stat 10:641–655
9. Kushner HJ, Yin GG (1997) Stochastic approximations algorithms and applications. Springer, New York
10. Lai TL (2003) Stochastic approximation. Ann Statis 31:391–406
11. Lin X, Breslow NE (1996) Bias correction in generalized linear mixed models with multiple components of dispersion. J Am Stat Assoc 91:1007–1016
12. Louis TA (1982) Finding the observed information matrix when using the em algorithm. J R Stat Soc Series B 44:226–233
13. McCulloch C (1997) Maximum likelihood algorithms for generalized linear mixed models. J Am Stat Assoc 92:162–170
14. Satten GA (1996) Rank-based inference in the proportional hazards model for interval censored data. Biometrika 83:355–370
15. Zhu HT, Lee SY (2003) Analysis of generalized linear mixed models via a stochastic approximation algorithm with markov chain Monte–Carlo method. Stat Comput 391–406

Chapter 8
Stochastic Comparisons of Spacings from Heterogeneous Samples

Subhash Kochar and Maochao Xu

Abstract In this paper we review some of the recently obtained results in the area of stochastic comparisons of sample spacings when the observations are not necessarily identically distributed. A few new results on necessary and sufficient conditions for various stochastic orderings among spacings are also given. The paper is concluded with some examples and applications.

8.1 Introduction

Spacings are of great interest in many areas of statistics, in particular, in the characterizations of distributions, goodness-of-fit tests, auction theory, life testing and reliability models. A large number of goodness-of-fit tests are based on functions of sample spacings (see [2, 14, 15]).

Let $X_1, \ldots, X_n$ be n nonnegative random variables. The random variables $D_{i:n} = X_{i:n} - X_{i-1:n}$ and $D^*_{i:n} = (n-i+1)D_{i:n}$, $i = 1, \ldots, n$, with $X_{0:n} \equiv 0$, are respectively called simple spacings and normalized spacings. In the reliability context they correspond to times elapsed between successive failures of components in a system. In stochastic auction theory, $D_{n:n}$ and $D_{2:n}$ are of particular interest, which represent auction rents in buyer's auction and reverse auction in the second-price business auction (see [37]). It is well known that the normalized spacings of a random sample from an exponential distribution are independent and identically distributed (i.i.d.) random variables having the same exponential distribution. Such a characterization may not hold for other distributions or when the observations are not independent and identically distributed.

In many cases, the observations are independent but not identically distributed and we call them as heterogeneous samples. The study of heterogeneous samples is of great interest in many areas. For examples, in engineering, a complex engineering

S. Kochar (✉)
Department of Mathematics and Statistics, Portland State University, Portland, OR 97201, USA
e-mail: kochar@pdx.edu

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_8,

system is often composed of many different types of electrical components. Investigating the reliability of such system is relying on heterogeneous samples, which are the failure times of electrical components collected from experiments. Accordingly, the topic of heterogenous samples plays an important role. However, such a study is often challenging as the distribution theory of spacings when the observations are not i.i.d. is often complicated. A powerful tool to investigate the stochastic properties of spacings of heterogeneous samples is *stochastic orders*, which is a widely studied concept in probability and statistics.

In this paper, we review some recent results on stochastic orderings between spacings of heterogeneous samples, and some new results and applications are presented as well. The other interesting topic not discussed in our paper is that of dependence among spacings. Interested readers may refer to recent papers by [4, 5, 13, 16, 30] and references therein on this topic.

8.2 Stochastic Orders

Let X and Y be two random variables with distributions F and G, and survival functions $\overline{F} = 1 - F$ and $\overline{G} = 1 - G$. If X is less likely than Y to take on large values, then, intuitively, the survival function of X will be smaller than the survival function of Y at any fixed point. This leads to the *usual stochastic order*.

Definition 8.1. X is said to be smaller than Y in the usual stochastic order, denoted by $X \leq_{st} Y$, if $\overline{F}(x) \leq \overline{G}(x)$, or equivalently, $F(x) \geq G(x)$.

Suppose that X and Y are life lengths of two electronic components and satisfy $X \leq_{st} Y$. If both components are observed to be alive at time $t > 0$, one might conjecture that the residual lives would also be stochastically ordered. However, such a result does not hold (cf. [24, 26])! Hence, a stronger concept than usual stochastic order is needed. The following order is motivated by the fact that the above conjecture is false.

Definition 8.2. X is said to be smaller than Y in the hazard rate order, denoted by $X \leq_{hr} Y$, if

$$P(X > x + t | X > t) \leq P(Y > x + t | Y > t), \quad \text{for all } x \geq 0 \text{ and all } t.$$

Now, suppose that two components X and Y failed before observation time $t > 0$. If $X \leq_{st} Y$, is it necessarily true that the 'inactive time' $[t - X | X \leq t]$ of X is stochastically larger that the 'inactive time' $[t - Y | Y \leq t]$ of Y? The answer is negative. The following definition is proposed to resolve this question.

Definition 8.3. X is said to be smaller than Y in the reverse hazard rate order, denoted by $X \leq_{rh} Y$, if

$$P(t - X > x | X \leq t) \geq P(t - Y > x | Y \leq t), \quad \text{for all } x \geq 0 \text{ and all } t.$$

An interesting order based on the mean residual life is defined as follows.

Definition 8.4. X is said to be smaller than Y in the mean residual life order, denoted by $X \leq_{mrl} Y$, if

$$\mathrm{E}(X_t) \leq \mathrm{E}(Y_t) \quad \text{for all } t.$$

There are other situations, where one would like to have

$$[X|X \in E] \leq_{st} [Y|Y \in E]$$

for all possible events E. It turns out that this requirement leads to the likelihood ratio order as defined below.

Definition 8.5. X is said to be smaller than Y in the likelihood ratio order, denoted by $X \leq_{lr} Y$, if

$$P(X > x|a < X \leq b) \leq P(Y > x|a < Y \leq b) \quad \text{for all } x \text{ and } a < b.$$

When the density functions of X and Y exist, $X \leq_{lr} Y$ is equivalent to

$$f(t)g(s) \leq f(s)g(t) \quad \text{for all } s \leq t.$$

The following chain of implications holds among the above univariate stochastic orders:

$$\begin{array}{ccccc} X \leq_{lr} Y & \Longrightarrow & X \leq_{hr} Y & \Longrightarrow & X \leq_{mrl} Y \\ \Downarrow & & \Downarrow & & \Downarrow \\ X \leq_{rh} Y & \Longrightarrow & X \leq_{st} Y & \Longrightarrow & \mathrm{E}(X) \leq \mathrm{E}(Y). \end{array} \tag{8.1}$$

An other order closely related to the likelihood ratio order is the joint likelihood ordering introduced by [34].

Definition 8.6. For a bivariate random variable (X, Y), X is said to be smaller than Y according to *joint likelihood ordering*, denoted by $X \overset{\ell r:j}{\leq} Y$, if and only if

$$E[\phi(X, Y)] \geq E[\phi(Y, X)], \quad \phi \in \mathcal{G}_{\ell r},$$

where

$$\mathcal{G}_{\ell r} : \{\phi : \phi(y, x) \leq \phi(x, y), \quad x \leq y\}.$$

It can be seen that

$$X_1 \overset{\ell r:j}{\leq} X_2 \;\Leftrightarrow\; f \in \mathcal{G}_{\ell r},$$

where $f(\cdot, \cdot)$ denotes the joint density of (X_1, X_2).

As pointed out by [34], joint likelihood ratio ordering between the components of a bivariate random vector may not imply likelihood ratio ordering between their marginal distributions unless the random variables are independent, but it does imply stochastic ordering between them, that is,

$$X \overset{\ell r:j}{\leq} Y \Rightarrow X \overset{st}{\leq} Y.$$

A bivariate function $\phi \in \mathcal{G}_{\ell r}$ is called *arrangement increasing* (AI). Hollander et al. [11] have studied many interesting properties of such functions, though, apparently, they did not relate it to the notion of likelihood ratio ordering.

The above idea can be extended to compare the components of an n-dimensional vector $\mathbf{X} = (X_1, \ldots, X_n)$. We define $X_1 \overset{\ell r:j}{\leq} \cdots \overset{\ell r:j}{\leq} X_n$ if the joint density $f(x_1, \ldots, x_n)$ of $\mathbf{X}$ is an *arrangement increasing function*. (See, [24] for the definition of an arrangement increasing function on $\mathcal{R}^n$.)

In a different context, [31] studied a subclass of arrangement increasing functions on $\mathcal{R}^n$, which they call as ISO^* functions, as described below. Let $\mathbf{x}$ and $\mathbf{y}$ be two vectors on $\mathcal{R}^n$ such that $\sum_{i=1}^{j} y_i \leq \sum_{i=1}^{j} x_i \;\; j = 1, \ldots, n-1$ and $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i$. We shall denote this partial ordering between the vectors by $\mathbf{x} \overset{\star}{\ll} \mathbf{y}$.

Definition 8.7. A real-valued function ϕ defined on a set $\mathcal{A} \subset \mathcal{R}^n$ is said to be ISO^* on $\mathcal{A}$ if $\phi(\mathbf{x}) \leq \phi(\mathbf{y}), \;\; \forall \;\; \mathbf{x} \overset{\star}{\ll} \mathbf{y}$.

As mentioned earlier, an ISO^* function is arrangement increasing but the converse is not true. It is easy to see that the joint density $f(x_1, x_2)$ of a bivariate random vector (X_1, X_2) is ISO^* if and only if the conditional density of X_2 given $X_1 + X_2 = t$ is monotonically increasing for each fixed t.

The usual likelihood ratio order has the following multivariate version.

Definition 8.8. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ be two n-dimensional random vectors with absolutely continuous [or discrete] distribution functions and let f and g denote their continuous [or discrete] density functions, respectively. Suppose that

$$f(\mathbf{x})g(\mathbf{y}) \leq f(\mathbf{x} \wedge \mathbf{y})g(\mathbf{x} \vee \mathbf{y})$$

for every $\mathbf{x}$ and $\mathbf{y}$. Then X is said to be smaller than Y in the multivariate likelihood ratio order (denoted as $\mathbf{X} \leq_{lr} \mathbf{Y}$).

It is known that multivariate likelihood ratio order implies component-wise likelihood ratio order. For more details on the univariate and multivariate stochastic orders (see [26, 33]).

A basic concept for comparing variability among distributions is that of dispersive ordering as defined below.

Definition 8.9. X is said to be less dispersed than Y, denoted by $X \leq_{disp} Y$, if

$$F^{-1}(\beta) - F^{-1}(\alpha) \leq G^{-1}(\beta) - G^{-1}(\alpha)$$

for all $0 < \alpha \leq \beta < 1$, where F^{-1} and G^{-1} denote their corresponding right continuous inverses.

For more discussion, one may refer to Sect. 3.B of [33].

A weaker order called the right spread order in [9], or *excess wealth* order in [32] has also been proposed to compare the variabilities of two distributions.

Definition 8.10. X is said to be less right spread than Y, denoted by $X \leq_{RS} Y$, if

$$\int_{F^{-1}(p)}^{\infty} \overline{F}(x)\, dx \leq \int_{G^{-1}(p)}^{\infty} \overline{G}(x) dx, \quad \text{for all } 0 \leq p \leq 1.$$

It is known that

$$X \leq_{disp} Y \Longrightarrow X \leq_{RS} Y \Longrightarrow \mathrm{Var}(X) \leq \mathrm{Var}(Y).$$

8.3 Spacings

As mentioned earlier, the normalized spacings of a random sample from an exponential distribution are i.i.d. random variables having the same exponential distribution. Many authors have studied the stochastic properties of spacings from restricted families of distributions. Barlow and Proschan [3] proved that if $X_1, \ldots, X_n$ is a random sample from a decreasing failure rate (DFR) distribution, then the successive normalized spacings are stochastically increasing. Kochar and Kirmani [20] strengthened this result from stochastic ordering to hazard rate ordering, that is, for $i = 1, \ldots, n-1$,

$$D^*_{i:n} \leq_{hr} D^*_{i+1:n}. \tag{8.2}$$

The corresponding problem when the random variables are not identically distributed has also been studied by many researchers. For a review of this topic see [19]. In the following, we review some important results obtained recently on this topic and also give some new results.

First, let us introduce the proportional hazard rates (PHR) model, which will be discussed extensively in the paper. Independent random variables $X_1, X_2, \ldots, X_n$ are said to follow PHR model if for $i = 1, 2, \ldots, n$, the survival function of X_i can be expressed as,

$$\overline{F}_i(x) = [\overline{F}(x)]^{\lambda_i}, \quad \text{for } \lambda_i > 0, \tag{8.3}$$

where $\overline{F}(x)$ is the survival function of some random variable X. If $r(t)$ denotes the hazard rate corresponding to the baseline distribution F, then the hazard rate of X_i is $\lambda_i r(t)$, $i = 1, 2, \ldots, n$. We can equivalently express (8.3) as

$$\overline{F}_i(x) = e^{-\lambda_i R(x)}, \quad i = 1, 2, \ldots, n$$

where $R(x) = \int_{-\infty}^{x} r(t)dt$, is the cumulative hazard rate of X. The PHR model includes many well known distributions, such as exponential, Weibull, Pareto and Lomax, etc.

8.3.1 One-Sample Problem

Pledger and Proschan [29] considered the problem of stochastically comparing the spacings of heterogeneous samples in the PHR model.

Theorem 8.1. *If $X_1, \ldots, X_n$ are independent random variables with X_i having survival function $\overline{F}(t) = e^{-\lambda_i R(t)}$, $i = 1, \ldots, n$, where $R(t)$ is a concave (i.e., F is DFR) and differentiable hazard function, then*

$$D^*_{i:n} \leq_{st} D^*_{j:n} \quad \text{for} \quad i \leq j.$$

Kochar and Kirmani [20] proved that the joint density of the normalized spacings is *ISO** when the joint density of the parent observations is convex. This will hold, in particular, when the X_i's are independent (but not necessarily identical) with log-convex densities.

Theorem 8.2. *Let the joint density $f_{\mathbf{X}}(x_1, \ldots, x_n)$ of $\mathbf{X} = (X_1, \ldots, X_n)$ be convex. Then the joint density of $\mathbf{D} = (D_{1:n}, \ldots, D_{n:n})$ is ISO**.

This result immediately leads to the following interesting result.

Theorem 8.3. *Let $X_1, X_2, \ldots, X_n$ be independent random variables with log-convex densities. Then*

$$D^*_{1:n} \overset{\ell r:j}{\leq} D^*_{2:n} \overset{\ell r:j}{\leq} \ldots \overset{\ell r:j}{\leq} D^*_{n:n}.$$

If a density is log-convex, it is DFR, but the converse is not true. Hence, Theorem 8.3 establishes a stronger ordering between the normalized spacings than does Theorem 8.1 under a stronger condition on the parent distributions. A related paper on this topic is by [27].

One may wonder whether under the condition of Theorem 8.3 the result can be extended from joint likelihood ratio order to likelihood ratio order? The answer is negative as illustrated by a counterexample in [25].

Let $X_1, \ldots, X_n$ be independent exponential random variables with X_i having failure rate λ_i for $i = 1, \ldots, n$. Hu et al. [12] proved the following result.

Theorem 8.4. *If $\lambda_{n+1} \leq [\geq] \lambda_k$ for $k = 1, \ldots, n$, then*

$$D_{n:n} \leq_{\text{lr}} D_{n+1:n+1}$$

and

$$D_{1:n} \leq_{\text{lr}} D_{2:n+1} [D_{2:n+1} \leq_{\text{lr}} D_{2:n}].$$

If $\lambda_i + \lambda_j \geq \lambda_k$ for all distinct i, j and k then

$$D_{n-1:n} \leq_{\text{lr}} D_{n:n}, \quad D_{n:n+1} \leq_{\text{lr}} D_{n:n}.$$

They also showed that $D_{2:3} \leq_{lr} D_{3:3}$ for all λ_i's, and

$$D_{1:n} \leq_{lr} D_{2:n}, \quad n \geq 2.$$

Observing (8.2), [21] posed the following interesting question. Let $X_1, \ldots, X_n$ be independent exponential random variables with possibly different scale parameters. Will it hold that

$$D^*_{i:n} \leq_{hr} D^*_{i+1:n}, \quad i = 1, \ldots, n-1? \tag{8.4}$$

They proved their conjecture for $n = 3$, but the general case still remains an open problem. This topic has been extensively studied by [6, 18, 35, 36] in single-outlier or multiple-outlier exponential models.

8.3.2 Two-Samples Problem

Pledger and Proschan [29] considered the problem of stochastically comparing the spacings of nonidentical independent exponential random variables with those corresponding to stochastically comparable independent and identically distributed exponential random variables.

Theorem 8.5. *If $X_1, \ldots, X_n$ are independent exponential random variables with X_i having hazard rate λ_i, $i = 1, \ldots, n$, and $Y_1, \ldots, Y_n$ is a random sample of size n from an exponential distribution with common hazard rate $\overline{\lambda}\lambda = \sum_{i=1}^n \lambda_i/n$, then*

$$H^*_{i:n} \leq_{st} D^*_{i:n}$$

where $H^*_{i:n} = (n-i+1)(Y_{i:n} - Y_{i-1:n})$ for $i = 1, \ldots, n$ are the normalized spacings of Y_i's, and $Y_{0:n} \equiv 0$.

Kochar and Korwar [21] strengthened and extended this result from stochastic ordering to likelihood ratio ordering and dispersive ordering.

Theorem 8.6. *If $X_1, \ldots, X_n$ are independent exponential random variables with X_i having hazard rate λ_i, $i = 1, \ldots, n$, and $Y_1, \ldots, Y_n$ is a random sample of size n from an exponential distribution with common hazard rate $\overline{\lambda}\lambda = \sum_{i=1}^n \lambda_i/n$, then*

$$H^*_{i:n} \leq_{lr} D^*_{i:n} \quad \text{and} \quad H^*_{i:n} \leq_{disp} D^*_{i:n}, \quad i = 1, \ldots, n.$$

Kochar and Rojo [22] further strengthened Theorem 8.6 to multivariate likelihood ratio order.

Theorem 8.7. *If $X_1, \ldots, X_n$ are independent exponential random variables with X_i having hazard rate λ_i, $i = 1, \ldots, n$, and $Y_1, \ldots, Y_n$ is a random sample of size n from an exponential distribution with common hazard rate $\overline{\lambda}\lambda = \sum_{i=1}^n \lambda_i/n$, then,*

$$(H^*_{1:n}, \ldots, H^*_{n:n}) \leq_{lr} (D^*_{1:n}, \ldots, D^*_{n:n})$$

As a consequence, [22] pointed out that

$$Y_{j:n} - Y_{i:n} \leq_{st} X_{j:n} - X_{i:n} \quad \text{for all} \quad 1 \leq i < j \leq n.$$

Let $X_1, \ldots, X_n$ be independent exponential random variables with X_i having hazard rate λ_i, $i = 1, \ldots, n$, and $Y_1, \ldots, Y_n$ be a random sample of size n from an exponential distribution with common hazard rate λ. One natural question is to find sufficient and necessary conditions for comparing simple sample spacings of $D_{k:n}$ from X_i's and $H_{k:n}$ from Y_i's. Kochar and Korwar [21] proved that the distribution function of $D^*_{k:n}$ for $i = 2, \ldots, n$ is a mixture of independent exponential random variables with the density function

$$f_{D^*_{k:n}}(x) = \sum_{\mathbf{r}} \frac{\prod_{i=1}^n \lambda_i}{\prod_{i=1}^n \sum_{j=i}^n \lambda_{r_j}} \left(\frac{\sum_{j=k}^n \lambda_{r_j}}{n-i+1} \right) \exp\left\{ \frac{-x \sum_{j=k}^n \lambda_{r_j}}{n-i+1} \right\},$$

where $\mathbf{r}$ extends over all of the permutations of $\{1, 2, \ldots, n\}$. Hence, the distribution of $D_{k:n}$ could be represented as

$$F_{D_{k:n}}(x) = \sum_{j \in \mathbf{r}} p_j F_{\lambda^*_{kj}}(x),$$

where j denotes a permutation of $(\lambda_1, \ldots, \lambda_n)$ belonging to $\mathbf{r}$ and

$$p_j = \frac{\prod_{i=1}^n \lambda_i}{\prod_{i=1}^n \sum_{j=i}^n \lambda_{r_j}},$$

and

$$\lambda^*_{kj} = \frac{\sum_{j=k}^n \lambda_{r_j}}{(n-k+1)^2},$$

and $F_{\lambda^*_{kj}}$ means an exponential distribution with hazard rate λ^*_{kj}.

Note that, the distribution function of $H_{k:n}$ is

$$F_{H_{k:n}}(x) = F_{\lambda/(n-k+1)}(x) = 1 - \exp\left\{ -\frac{\lambda x}{n-k+1} \right\}.$$

The following results provide sufficient and necessary conditions for stochastically comparing $D_{k:n}$ and $H_{k:n}$ according to likelihood ratio and reverse hazard rate orderings.

Theorem 8.8. *Let $X_1, \ldots, X_n$ be independent exponential random variables with X_i having hazard rate λ_i, $i = 1, \ldots, n$, and $Y_1, \ldots, Y_n$ be a random sample of size n from an exponential distribution with common hazard rate λ. Then, for $k \geq 2$,*

(a) $H_{k:n} \leq_{lr} D_{k:n}$ or

(b) $H_{k:n} \leq_{rh} D_{k:n}$

if and only if

$$\lambda \geq \frac{1}{n-k+1} \frac{\sum_{\mathbf{r}} \frac{\prod_{i=1}^{n} \lambda_i}{\prod_{i=1}^{n} \sum_{j=i}^{n} \lambda_{r_j}} \left(\sum_{j=k}^{n} \lambda_{r_j}\right)^2}{\sum_{\mathbf{r}} \frac{\prod_{i=1}^{n} \lambda_i}{\prod_{i=1}^{n} \sum_{j=i}^{n} \lambda_{r_j}} \sum_{j=k}^{n} \lambda_{r_j}}.$$

Proof. The necessary and sufficient condition for likelihood ratio order between $H_{k:n}$ and $D_{k:n}$ follows from Lemma 3.1 of [40]. It remains to prove that the condition on λ is a necessary condition for the reverse hazard rate order.

Note that the reverse hazard rate of $D_{k:n}$ for $k \geq 2$, by Taylor's expansion around zero, is

$$\begin{aligned}
\tilde{r}_{D_{k:n}}(x) &= \frac{\sum_{j\in\mathbf{r}} p_j \lambda_{kj}^* \exp\{-\lambda_{kj}^* x\}}{\sum_{j\in\mathbf{r}} p_j (1-\exp\{-\lambda_{kj}^* x\})} \\
&= \frac{\sum_{j\in\mathbf{r}} p_j \lambda_{kj}^* (1-\lambda_{kj}^* x + o(x))}{\sum_{j\in\mathbf{r}} p_j \lambda_{kj}^* x + o(x)} \\
&= \frac{\sum_{j\in\mathbf{r}} p_j \lambda_{kj}^* - \sum_{j\in\mathbf{r}} p_j (\lambda_{kj}^*)^2 x + o(x)}{\sum_{j\in\mathbf{r}} p_j \lambda_{kj}^* x + o(x)}.
\end{aligned}$$

Similarly, for $i \geq 2$,

$$\tilde{r}_{H_{k:n}}(x) = \frac{\lambda - \lambda^2 x/(n-k+1) + o(x)}{\lambda x + o(x)}.$$

Then

$$\tilde{r}_{H_{k:n}}(x) \leq \tilde{r}_{D_{k:n}}(x)$$

implies

$$\lambda \geq (n-k+1) \frac{\sum_{j\in\mathbf{r}} p_j \left(\lambda_{kj}^*\right)^2}{\sum_{j\in\mathbf{r}} p_j \lambda_{kj}^*}.$$

The result follows immediately. □

Analogously, we have an equivalent necessary and sufficient condition for hazard rate order, dispersive order and stochastic order between $D_{k:n}$ and $H_{k:n}$.

Theorem 8.9. *Let $X_1, \ldots, X_n$ be independent exponential random variables with X_i having hazard rate λ_i, $i = 1, \ldots, n$, and $Y_1, \ldots, Y_n$ be a random sample of size n from an exponential distribution with common hazard rate λ. Then, for $k \geq 2$,*

(a) $H_{k:n} \leq_{hr} D_{k:n}$ *or*
(b) $H_{k:n} \leq_{disp} D_{k:n}$ *or*
(c) $H_{k:n} \leq_{st} D_{k:n}$

if and only if

$$\lambda \geq \frac{1}{n-k+1} \sum_{\mathbf{r}} \frac{\prod_{i=1}^{n} \lambda_i}{\prod_{i=1}^{n} \sum_{j=i}^{n} \lambda_{r_j}} \sum_{j=k}^{n} \lambda_{r_j}.$$

Proof. Using Lemma 2.1 in [28], it is easy to show the condition on λ is a necessary and sufficient condition for hazard rate order between $D_{k:n}$ and $H_{k:n}$. Since $H_{k:n}$ is an exponential random variable, according to Theorem 3.B.20 of [33], (a) implies (b). By Theorem 3.B.13 there, (b) implies (c). Hence, it is enough to show the condition on λ is the necessary condition for the stochastic order. Using Taylor's expansion around zero for the distribution function, for $k \geq 2$,

$$F_{D_{k:n}}(x) = \sum_{j \in \mathbf{r}} p_j (\lambda_{kj}^* x + o(x)),$$

and

$$F_{H_{k:n}}(x) = \frac{\lambda}{n-k+1} x + o(x).$$

So,

$$F_{H_{k:n}}(x) \geq F_{D_{k:n}}(x)$$

implies,

$$\lambda \geq (n-k+1) \sum_{j \in \mathbf{r}} p_j \lambda_{kj}^*.$$

Hence, the required result follows. □

Below is another result in the same direction.

Theorem 8.10. *Let $X_1, \ldots, X_n$ be independent exponential random variables with X_i having hazard rate λ_i, $i = 1, \ldots, n$, and $Y_1, \ldots, Y_n$ be a random sample of size n from an exponential distribution with common hazard rate λ. Then, for $k \geq 2$,*

(a) $H_{k:n} \leq_{mrl} D_{k:n}$ *or*
(b) $H_{k:n} \leq_{RS} D_{k:n}$ *or*
(c) $\mathrm{E}H_{k:n} \leq \mathrm{E}D_{k:n}$

if and only if

$$\lambda \geq \frac{1}{n-k+1} \frac{1}{\sum_{\mathbf{r}} \frac{\prod_{i=1}^{n} \lambda_i}{\sum_{j=k}^{n} \lambda_{r_j} \prod_{i=1}^{n} \sum_{j=i}^{n} \lambda_{r_j}}}.$$

Proof. From Lemma 2.6 of [38], it is seen that the condition on λ is necessary and sufficient for the mean residual life order. By Theorem 3.C.6 of [33], (a) implies (b). It is easy to see that (b) implies (c). Hence, it is enough to show the condition on λ is a necessary condition for (c). Note that

$$\mathrm{E}(D_{k:n}) = \sum_{\mathbf{r}} \frac{p_j}{\lambda_{kj}^*},$$

and

$$\mathrm{E}(H_{k:n}) = \frac{n-k+1}{\lambda},$$

the result follows immediately. □

Example 8.1. Suppose X_1, X_2, X_3 are independent exponential random variables with parameters $\lambda_1, \lambda_2, \lambda_3$, respectively; and Y_1, Y_2, Y_3 are i.i.d. exponential random variables with parameter λ. Then, from Theorem 8.9,

$$Y_{3:3} - Y_{2:3} \leq_{hr} X_{3:3} - X_{2:3}$$

if and only if

$$\lambda \geq \lambda_{hr} = \frac{2\lambda_1\lambda_2\lambda_3}{\lambda_1+\lambda_2+\lambda_3}\left(\frac{1}{\lambda_2+\lambda_3} + \frac{1}{\lambda_1+\lambda_3} + \frac{1}{\lambda_1+\lambda_2}\right).$$

From Theorem 8.8,

$$Y_{3:3} - Y_{2:3} \leq_{lr} X_{3:3} - X_{2:3}$$

if and only if

$$\lambda \geq \lambda_{lr} = \frac{3}{2}\left(\frac{1}{\lambda_2+\lambda_3} + \frac{1}{\lambda_1+\lambda_3} + \frac{1}{\lambda_1+\lambda_2}\right)^{-1}.$$

From Theorem 8.10

$$Y_{3:3} - Y_{2:3} \leq_{RS} X_{3:3} - X_{2:3}$$

if and only if

$$\lambda \geq \lambda_{RS} = \left[\frac{\lambda_1\lambda_2\lambda_3}{\lambda_1+\lambda_2+\lambda_3}\left(\frac{\lambda_2^2+\lambda_3^2}{(\lambda_2+\lambda_3)\lambda_2^2\lambda_3^2} + \frac{\lambda_1^2+\lambda_3^2}{(\lambda_1+\lambda_3)\lambda_1^2\lambda_3^2} + \frac{\lambda_2^2+\lambda_1^2}{(\lambda_2+\lambda_1)\lambda_2^2\lambda_1^2}\right)\right]^{-1}.$$

It is worth noting that from (8.1), it follows that

$$\lambda_{lr} \geq \lambda_{hr} \geq \lambda_{RS}.$$

For example, let

$$\lambda_1 = 1, \lambda_2 = 2, \lambda_3 = 3.$$

Then,

$$\lambda_{lr} = 1.91489 > \lambda_{hr} = 1.56667 > \lambda_{RS} = 1.30435.$$

8.4 Sample Range

Sample range is one of the criteria for comparing variabilities among distributions and hence it is important to study its stochastic properties.

Kochar and Rojo [22] proved the following result.

Theorem 8.11. *If $X_1, \ldots, X_n$ are independent exponential random variables with X_i having hazard rate λ_i, $i = 1, \ldots, n$, and $Y_1, \ldots, Y_n$ is a random sample of size n from an exponential distribution with common hazard rate $\overline{\lambda} = \sum_{i=1}^{n} \lambda_i / n$, then,*

$$Y_{n:n} - Y_{1:n} \leq_{st} X_{n:n} - X_{1:n}.$$

The next theorem due to [17] improves upon this result.

Theorem 8.12. *If $X_1, \ldots, X_n$ are independent exponential random variables with X_i having hazard rate λ_i, $i = 1, \ldots, n$, and $Y_1, \ldots, Y_n$ is a random sample of size n from an exponential distribution with common hazard rate $\tilde{\lambda} = \left(\prod_{i=1}^{n} \lambda_i\right)^{1/n}$, then,*

$$Y_{n:n} - Y_{1:n} \leq_{st} X_{n:n} - X_{1:n}.$$

Recently, [39] obtained a necessary and sufficient condition for stochastically comparing sample ranges from heterogeneous and homogeneous exponential samples.

Theorem 8.13. *If $X_1, \ldots, X_n$ are independent exponential random variables with X_i having hazard rate λ_i, $i = 1, \ldots, n$, and $Y_1, \ldots, Y_n$ is a random sample of size n from an exponential distribution with common hazard rate λ, then,*

$$Y_{n:n} - Y_{1:n} \leq_{st} X_{n:n} - X_{1:n} \Longleftrightarrow \lambda \geq \hat{\lambda} = \left(\frac{\prod_{i=1}^{n} \lambda_i}{\overline{\lambda}}\right)^{1/(n-1)},$$

where $\overline{\lambda} = \sum_{i=1}^{n} \lambda_i / n$.

Note that $\hat{\lambda} \leq \tilde{\lambda}$, which improves upon the results of Theorem 8.12.

Kochar and Xu [23] partially strengthened Theorem 8.11 from stochastic order to reverse hazard rate order.

Theorem 8.14. *Let $X_1, \ldots, X_n$ be independent exponential random variables with X_i having hazard rate λ_i, $i = 1, \ldots, n$. Let $Y_1, \ldots, Y_n$ be a random sample of size n from an exponential distribution with common hazard rate $\overline{\lambda} = \sum_{i=1}^{n} \lambda_i / n$. Then*

$$Y_{n:n} - Y_{1:n} \leq_{rh} X_{n:n} - X_{1:n}.$$

They also proved the following theorem for the PHR model.

Theorem 8.15. *Let $X_1, \ldots, X_n$ be independent random variables with X_i having survival function $\overline{F}^{\lambda_i}$, $i = 1, \ldots, n$. Let $Y_1, \ldots, Y_n$ be a random sample with common population survival distribution $\overline{F}^{\overline{\lambda}}$, where $\overline{\lambda} = \sum_{i=1}^{n} \lambda_i / n$, then*

$$Y_{n:n} - Y_{1:n} \leq_{st} X_{n:n} - X_{1:n}.$$

More recently, [10] proved the following result.

Theorem 8.16. *Let $X_1, \ldots, X_n$ be independent exponential random variables with X_i having hazard rate λ_i, $i = 1, \ldots, n$, and $Y_1, \ldots, Y_n$ be a random sample of size n from an exponential distribution with common hazard rate $\overline{\lambda} = \sum_{i=1}^{n} \lambda_i / n$. Let $\tilde{r}_X(x)$ and $\tilde{r}_Y(x)$ be reversed hazard rates of $X_{n:n} - X_{1:n}$ and $Y_{n:n} - Y_{1:n}$, respectively. Then*

$$\frac{\tilde{r}_X(x)}{\tilde{r}_Y(x)} \quad \textit{is increasing.}$$

The following results follow as consequences as shown in that paper.

Corollary 8.1. *Let $X_1, \ldots, X_n$ be independent exponential random variables with X_i having hazard rate λ_i, $i = 1, \ldots, n$. Let $Y_1, \ldots, Y_n$ be a random sample of size n from an exponential distribution with common hazard rate $\overline{\lambda}$, then,*

$$Y_{n:n} - Y_{1:n} \leq_{lr} X_{n:n} - X_{1:n},$$

and

$$Y_{n:n} - Y_{1:n} \leq_{disp} X_{n:n} - X_{1:n}.$$

8.5 Applications

8.5.1 Type-II Censoring

If n items are put on life test and the test terminated at the time of rth failure, then

$$T_{rn} = \sum_{i=1}^{r} (n - i + 1)(X_{i:n} - X_{i-1:n}) = \sum_{i=1}^{r} D_{i:n}^{*}$$

represents the total time on test (TTT), which is a well-known concept in statistical reliability. As pointed out in [8], if those component lifetimes are i.i.d. exponential with mean θ or equivalently the failure rate $\lambda = 1/\theta$, the best estimator (UMVUE) of θ is

$$\hat{\theta} = \frac{T_{rn}}{r}.$$

It is well-known that (cf., [8]),

$$\frac{2T_{rn}}{r} \xrightarrow{L} \chi^2_{2r},$$

where χ^2_{2r} denotes a chi-square distribution with $2r$ degrees of freedom. Based on this result, many estimation and testing statistical procedures have been established. However, if the components are actually from different exponential distributions with failure rates $\lambda_1^*, \ldots, \lambda_n^*$, then it follows from Theorem 8.7 and Theorem 6.E.8 of [33] that

$$\hat{\theta}^* \geq_{st} \hat{\theta},$$

where

$$\hat{\theta}^* = \frac{\sum_{i=1}^{r} D^*_{i:n}}{r}$$

and $\lambda = \sum_{i=1}^{n} \lambda_i^*/n$. This means that based on TTT statistic, one would overestimate the average lifetime of components using heterogeneous exponential components.

8.5.2 Reliability

An n-component system that fails if and only if at least k of the n components fail is called a k-out-of-n: F system. The lifetime of such a system could be represented as $X_{k:n}$. The k-out-of-n system structure is a very popular type of redundancy in fault tolerant systems with wide applications in industrial and military systems. For two different systems say a k-out-of-n: F system and a $k+1$-out-of-n: F system, the engineer may be interested in the additional lifetime $X_{k+1:n} - X_{k:n}$ for the system design and the cost purpose. Due to the complicated distribution form, one may provide a sharp bound on the survival function of this based on Theorem 8.9.

For example, a plane has four engines, and a minimum of three engines are required for the plane work. Hence, this plane is a 2-out-of-4: F system. If the engineer is planning to improve the system to the 3-out-of-4: F system, he/she has to consider the cost and the reliability of improvement simultaneously. Hence, it is important to estimate the survival probability and the mean additional lifetime $X_{3:4} - X_{2:4}$. Theorem 8.9 provides a lower bound for the survival function of the additional lifetime. Let $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (1, 2, 2.5, 3)$, then,

$$P(X_{3:4} - X_{2:4} > x) \geq e^{-0.955x}.$$

Using Theorem 8.10, the following sharp lower bound could be established,

$$\mathrm{E}(X_{3:4} - X_{2:4}) \geq 1.0879.$$

In the following table, we list the lower bounds on λ's for various partial orders to hold in Theorems 8.8–8.10.

λ	lr/ rh	hr/ disp/ st	mrl/ RS/ E
$D_{2:4}$	2.060897	2.039216	2.019271
$D_{3:4}$	1.988454	1.909626	1.838388
$D_{4:4}$	2.009722	1.674501	1.382051

The values above are comparable to the arithmetic mean 2.125, geometric mean 1.96799 and harmonic mean 1.791045 of λ_i's. It is interesting to note that all of the values are less than the arithmetic mean, which coincides with the condition in Theorem 8.7.

8.5.3 Dependence Orderings Among Order Statistics

Recently [1] proved that, in a very general sense, the dependence between pairs of order statistics decreases as the indices of the order statistics draw apart when the observations are independent and identically distributed. Their proofs make use of the various results on stochastic orderings between spacings and order statistics in a nice way.

Let $X_1, \ldots, X_n$ be independent continuous random variables with X_i having survival function $\overline{F}^{\lambda_i}$, $i = 1, \ldots, n$, and $Y_1, \ldots, Y_n$ be i.i.d. continuous random variables. Dolati et al. [7] used the concept of more RTI (right-tail increasing) dependence order to investigate the relative dependence between the extreme order statistics. They showed that

$$(X_{n:n}|X_{1:n}) \prec_{\text{RTI}} (Y_{n:n}|Y_{1:n}).$$

They wondered whether the result can be extended from more RTI dependence order to more SI (stochastic increasing) dependence order which is a stronger dependence order. Genest et al. [10] have successively made use of Corollary 8.1 to solve this problem by proving

$$(X_{n:n}|X_{1:n}) \prec_{\text{SI}} (Y_{n:n}|Y_{1:n}).$$

References

1. Avérous J, Genest C, Kochar SC (2005) On the dependence structure of order statistics. J Multivar Anal 94:159–171
2. Balakrishnan N, Rao CR (1998) Handbook of statistics 16-order statistics: Theory and methods. Elsevier, New York
3. Barlow RE, Proschan F (1966) Inequalities for linear combinations of order statistics from restricted families. Ann Math Statist 37:1574–1592

4. Bon JL, Păltănea E (2006) Comparison of order statistics in a random sequence to the same statistics with i.i.d. variables. ESAIM Probab Stat 10:1–10
5. Burkschat M (2009) Multivariate dependence of spacings of generalized order statistics. J Multivar Anal 100:1093–1106
6. Chen H, Hu T (2008) Multivariate likelihood ratio orderings between spacings of heterogeneous exponential random variables. Metrika 68:17–29
7. Dolati A, Genest C, Kochar SC (2008) On the dependence between the extreme order statistics in the proportional hazards model. J Multivar Anal 99:777–786
8. Epstein B (1960) Estimation from life test data. Technometric 2:447–454
9. Fernández-Ponce JM, Kochar SC, Muñoz-Perez J (1998) Partial orderings of distributions based on right-spread functions. J Appl Probab 35:221–228
10. Genest C, Kochar S, Xu M (2009) On the range of heterogeneous samples. J Multivar Anal 100:1587–1592
11. Hollander M, Proschan F, Sethuraman J (1977) Functions decreasing in transposition and their applications in ranking problems. Ann Stat 4:722–733
12. Hu T, Lu Q, Wen S (2008) Some new results on likelihood ratio orderings for spacings of heterogeneous exponential random variables. Commun Stat Theory Methods 37:2506–2515
13. Hu T, Wang F, Zhu Z (2006) Stochastic comparisons and dependence of spacings from two samples of exponential random variables. Commun Stat Theory Methods 35:979–988
14. Jammalamadaka SR, Goria MN (2004) A test of goodness of git based on Gini's index of spacings. Stat Prob Lett 68:177–187
15. Jammalamadaka SR, Taufer E (2003) Testing exponentiality by comparing the empirical distribution function of the normalized spacings with that of the original data. J Nonparametr Stat 15:719–729
16. Khaledi B, Kochar SC (2000a) Dependence among spacings. Probab Eng Inform Sci 14: 461–472
17. Khaledi B, Kochar SC (2000b) Sample range-some stochastic comparisons results. Calcutta Stat Assoc Bull 50:283–291
18. Khaledi B, Kochar SC (2001) Stochastic properties of spacings in a single-outlier exponential model. Probab Eng Inform Sci 15:401–408
19. Kochar SC (1998) Stochastic comparisons of spacings and order statistics. In: Basu AP, Basu SK, Mukhopadhyay S (eds) Frontiers in reliability. World Scientific, Singapore, pp 201–216
20. Kochar SC, Kirmani S (1995) Some results on normalized spacings from restricted families of distributions. 46:47–57
21. Kochar SC, Korwar R (1996) Stochastic orders for spacings of heterogeneous exponential random variables. J Multivar Anal 59:272–281
22. Kochar SC, Rojo J (1996) Some new results on stochastic comparisons of spacings from heterogeneous exponential distributions. J Multivar Anal 57:69–83
23. Kochar SC, Xu M (2007) Stochastic comparisons of parallel systems when components have proportional hazard rates. Probab Eng Inform Sci 21:597–609
24. Marshall AW, Olkin I (2007) Life distributions. Springer, New York
25. Misra N, van der Meulen EC (2003) On stochastic properties of m-spacings. J Stat Plan Inference 115:683–697
26. Müller A, Stoyan D (2002) Comparison methods for stochastic models and risks. Wiley, New York
27. Nappo G, Spizzichino F (1998) Ordering properties of the TTT-plot of lifetimes with Schur joint densities. Stat Prob Lett 39:195–203
28. Păltănea E (2008) On the comparison in hazard rate ordering of fail-safe systems. J Stat Plan Inference 138:1993–1997
29. Pledger P, Proschan F (1971) Comparisons of order statistics and of spacings from heterogeneous distributions. In: Rustagi JS (ed) Optimizing methods in statistics. Academic, New York, pp 89–113
30. Proschan F, Sethuraman J (1976) Stochastic comparisons of order statistics from heterogeneous populations, with applications in reliability. J Multivariate Anal 6:608–616

31. Robertson T, Wright FT (1982) On measuring the conformity of a parameter set to a trend, with applications. Ann Stat 4:1234–1245
32. Shaked M, Shantikumar JG (1998) Two variability orders. Probab Eng Inform Sci 12:1–23
33. Shaked M, Shanthikumar JG (2007) Stochastic orders and their applications. Springer, New York
34. Shanthikumar JG, Yao DD (1991) Bivariate characterization of some stochastic order relations. Adv Appl Prob 23:642–659
35. Wen S, Lu Q, Hu T (2007) Likelihood ratio orderings of spacings of heterogeneous exponential random variables. J Multivar Anal 98:743–756
36. Xu M, Li X, Zhao P, Li Z (2007) Likelihood ratio order of m-spacings in multipleoutlier models. Commun Stat Theory Methods 36:1507–1525
37. Xu M, Li X (2008) Some further results on the winner's rent in the second-price business auction. Sankhya 70:124–133
38. Zhao P, Balakrishnan N (2009) Characterization of MRL order of fail-safe systems with heterogeneous exponential components. J Stat Plan Inference, doi:10.1016/j.jspi.2009.02.006
39. Zhao P, Li X (2009) Stochastic order of sample range from heterogeneous exponential random variables. Probab Eng Inform Sci 23:17–29
40. Zhao P, Li X, Balakrishnan N (2009) Likelihood ratio order of the second order statistic from independent heterogeneous exponential random variables. J Multivar Anal 100:952–962

Chapter 9
The Distributions of the Peak to Average and Peak to Sum Ratios Under Exponentiality

Tomasz J. Kozubowski, Anna K. Panorska, and Fares Qeadan

Abstract Let $E_1, E_2, \ldots, E_N$ be independent and identically distributed exponential random variables, and let $Y = \bigvee_{i=1}^{N} E_i$ and $S = \sum_{i=1}^{N} E_i$ be their maximum and sum, respectively. We review distributional properties of the peak to sum and peak to average ratios, $R = Y/S$ and $\tilde{R} = Y/(S/N)$, respectively, with deterministic N, and provide an extension to the case where N is itself a random variable, independent of the $\{E_j\}$. Our results include explicit formulas for the relevant density and distribution functions, which apply to any distribution of N, as well as a particular example with geometrically distributed N. An example from climatology shows modeling potential of these models.

9.1 Introduction

Let $E_1, E_2, \ldots$ be independent and identically distributed (IID) exponential random variables with the probability density function (PDF)

$$f(x) = \beta e^{-\beta x}, \quad x > 0, \tag{9.1}$$

and let N be a discrete random variable taking values in the set $\mathbb{N} = \{1, 2, 3, \ldots\}$ of positive integers, independent of the $\{E_i\}$. Let $X = \sum_{i=1}^{N} E_i$ be the sum and $Y = \bigvee_{i=1}^{N} E_i$ be the maximum (peak) of the random number N of the $\{E_i\}$. In this paper we study distributions of the following two ratios: the peak to sum ratio

$$R = \frac{Y}{X} = \frac{\bigvee_{i=1}^{N} E_i}{\sum_{i=1}^{N} E_i}, \tag{9.2}$$

T.J. Kozubowski (✉)
Department of Mathematics and Statistics, University of Nevada, MS 084, Reno, NV 89557, USA
e-mail: tkozubow@unr.edu

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_9,

and the peak to average ratio

$$\tilde{R} = \frac{Y}{X/N} = \frac{\bigvee_{i=1}^{N} E_i}{\frac{1}{N}\sum_{i=1}^{N} E_i}. \tag{9.3}$$

While our results are valid for any distribution of N, we concentrate on the particular case where N is a geometric random variable $N \sim Geo(p)$ with the PDF

$$p_n = \mathbb{P}(N = n) = p(1-p)^{n-1}, \quad n \in \mathbb{N}. \tag{9.4}$$

Let us note that when the number of the $\{E_j\}$ is a random variable, there is no simple relation between the distributions of the two ratios R and $\tilde{R}$, which is in contrast with the case when N is deterministic.

The distributions of the rations have been studied for many years, because of their importance in various applications, especially in engineering, communications, computer engineering, scheduling, and insurance (see, e.g., [1, 6, 17, 21] and the references therein). Perhaps the most important applications of R and $\tilde{R}$ are in engineering, where *peak to average power ratio* typically needs to be minimized, see [21] and the references therein. However, very little is known about the exact distributions of the ratios with general distribution of the $\{E_j\}$ and either deterministic or random N, except for the case of exponential distribution and deterministic N, for which explicit formulas are available for the PDF as well as the CDF. These results, which actually go back to [8], appeared in the context of testing for outliers in exponential samples (see [12, 15]), where the ratios of order statistics to their sum are in common use. However, these results are not widely known. In one of the earliest works, the authors [17] derive the exact distribution of the peak to *median* ratio, arguing that using median rather than mean allows for standard derivation as both the peak (maximum) and the median are order statistics. There are also numerous works dealing with the asymptotic results for the ratios, as well as for the distribution of the vector (X, Y), none of which present the explicit exponential case (see, e.g., [2,4–6,9,10,16,18]). We hope that our brief summary of the deterministic case with exponentially distributed $\{E_j\}$ will be helpful in this regard.

We consider the ratios (9.2) and (9.3) with both deterministic and random N. We shall denote the distribution of R by PTSRn or PTSRN (peak to sum ratio), according to whether the number of terms is deterministic or random. Similarly, we shall use the notation PTARn or PTARN (peak to average ratio) for the distribution of $\tilde{R}$. We first review the exponential case with deterministic number of terms in Sect. 9.2. In Sect. 9.3 we present new general results for random N. A special case with geometric N is treated in Sect. 9.4, while a practical illustrative data example from climatology appears in Sect. 9.5. Technical derivations are deferred to the last section.

9.2 Models with Deterministic Number of Terms

Here we review basic results related to the ratios R and $\tilde{R}$ with deterministic number of terms, including PDF and CDF for both ratios.

9.2.1 Peak to Sum Ratio

Let R have a PTSRn distribution defined by (9.2), where $N = n$ is non-random and $E_1, E_2, \ldots E_n$ are IID with exponential distribution (9.1). Clearly, when $n = 1$, the distribution of R is a point mass at 1 with probability one. One way to obtain the PDF of $R = Y/X$ for general $n \geq 2$, is via the joint distribution of the sum $X = \sum_{j=1}^{n} E_j$ and the maximum $Y = \bigvee_{i=1}^{N} E_i$ of IID exponentials, recently derived in [20]. A standard transformation $R = Y/X$, $S = X$ leads to the joint PDF of (R, S), which actually turns out to be the product of the marginal PDFs of R and S, showing the independence of these quantities – which is a well-known property of exponential samples – as well as the exact form of the PDF of R, shown below. The support of the distribution of R is the interval $[1/n, 1]$, and for each $t \in [1/n, 1]$, the PDF of R is

$$h_n(t) = \begin{cases} \sum_{s=1}^{k} a_n(s)(1-st)^{n-2} & \frac{1}{k+1} \leq t \leq \frac{1}{k}, \;\; k = 1, 2, \ldots, n-1 \\ 0 & \text{otherwise,} \end{cases} \tag{9.5}$$

where

$$a_n(s) = n(n-1)\binom{n-1}{s-1}(-1)^{s+1}. \tag{9.6}$$

Subsequent integration of the PDF produces the CDF, which turns out to be

$$H_n(t) = \begin{cases} 0, & t < 1/n, \\ \sum_{s=1}^{k} \binom{n}{s}(-1)^s(1-st)^{n-1} + 1, & \frac{1}{k+1} \leq t < \frac{1}{k}, k = 1, 2, \ldots, n-1, \\ 1 & t \geq 1. \end{cases} \tag{9.7}$$

Remark 9.1. It is worth noting that the quantity $a_n(s)$ in (9.2) arises as the likelihood ratio test statistic for an upper outlier in exponential samples (see, e.g., [7, 11–13, 15]), and its distribution was derived in this context in [15]. This was accomplished by first establishing the recursive relation

$$h_n(t) = n f_{1,n-1}(t) H_{n-1}\left(\frac{t}{1-t}\right), \;\; t \in [1/n, 1], \tag{9.8}$$

between the PDF and the CDF of R, where

$$f_{a,b}(t) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1}(1-t)^{b-1}, \quad t \in [0,1], \tag{9.9}$$

is the PDF of a beta distribution with parameters $a, b \in (0,1)$. Repeated application of (9.8) leads to the formula (9.7), which was actually derived earlier in [8] via a rather heuristic arguments. Differentiation of (9.7) produces the PDF (9.5). Further, as noted in [15], similar considerations apply to the ratio connected with random samples from a gamma distribution with shape parameter $\alpha > 0$, in which case the recursive relation (9.8) still holds with $f_{1,n-1}$ replaced by $f_{\alpha,\alpha(n-1)}$.

Remark 9.2. Notice that the tail probability, $\mathbb{P}(R > t)$, takes a particularly simple form for $t \in (1/2, 1)$, since we have $\mathbb{P}(R > t) = n(1-t)^{n-1}$.

Remark 9.3. Observe that when $n = 2$, the ratio R has a uniform distribution on its supporting interval $[1/2, 1]$, while for $n = 3$ we get a triangular distribution on the interval $[1/3, 1]$ with a unique mode $m_3 = 1/2$ (see Fig. 9.1). It can be shown by elementary calculations that for $n \geq 4$ the densities (9.5) are differentiable on the interval $(1/n, 1)$, with the values of zero at the endpoints. It also appears that all densities are unimodal with the unique mode in the interval $[1/n, 1/2]$, which can be proven by straightforward algebra for the cases $n = 4$, and 5, where the modes are $m_4 = 5/11$ and $m_5 = (30 - \sqrt{32})/62$, respectively.

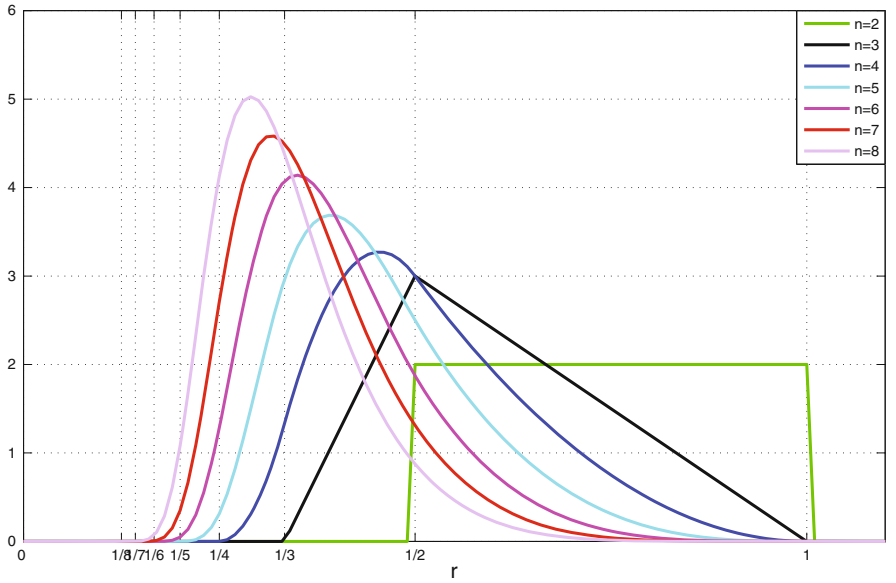


Fig. 9.1 The densities (9.5) of the ratio (9.2) for $n = 2, 3, \ldots, 8$

9.2.2 Peak to Average Ratio

The CDF and PDF of the peak to average ratio $\tilde{R}$ defined in (9.3) with $N = n$ (denoted by PTARn) are easily derived from those of PTSRn via the relation $\tilde{R} \stackrel{d}{=} nR$. The PDF of $\tilde{R}$ is given by $g_n(t) = (1/n)h_n(t/n)$, where $h_n(x)$ is the PDF of R defined in (9.5). Note, that this PDF is nonzero for $t \geq 1$. In turn, the CDF of PTARn is $G_n(t) = H_n(t/n)$, where $H_n(x)$ is the CDF of PTSRn given above.

9.3 Models with Random Number of Terms

In this section, we consider a general case, were N is an integer-valued random variable supported on the set $\mathbb{N}$, with

$$\mathbb{P}(N = n) = p_n, \quad n \in \mathbb{N}, \tag{9.10}$$

while in the following section we present a particular example where N has a geometric distribution.

9.3.1 Peak to Sum Ratio

To obtain the CDF of the PTSRN distribution, we use a standard conditioning argument. First, observe that the probability $\mathbb{P}(R \leq x)$ is zero whenever $x \leq 0$ and one if $x \geq 1$. Second, for any $0 < x < 1$, we have

$$\mathbb{P}(R \leq x) = \sum_{n=1}^{\infty} \mathbb{P}(R \leq x | N = n)\mathbb{P}(N = n) = \sum_{n=1}^{\infty} H_n(x)\, p_n,$$

with H_n and p_n given by (9.7) and (9.10), respectively. Assuming that $1/(k+1) \leq x < 1/k$ with some $k \in \mathbb{N}$ and taking into account (9.7), after some algebra we obtain

$$\mathbb{P}(R \leq x) = \sum_{s=1}^{k} (-1)^s \sum_{n=k+1}^{\infty} \binom{n}{s} (1 - sx)^{n-1} p_n + \mathbb{P}(N \geq k+1). \tag{9.11}$$

Note that the function (9.11) is continuous in x on the interval $(0, 1)$, with the one-sided limits of

$$\lim_{x \to 0^-} \mathbb{P}(R \leq x) = 0, \quad \lim_{x \to 1^+} \mathbb{P}(R \leq x) = 1 - p_1.$$

The CDF of R is continuous on $\mathbb{R}$ except at $x = 1$, where there is a jump of p_1. Thus, as expected, the distribution of peak to sum ratio R is a mixture of a point mass at 1 (with probability p_1) and a continuous distribution on the interval $(0, 1)$ (with probability $1 - p_1$), with the latter given by the CDF $F(t)$:

$$F(t) = \begin{cases} 0, & t < 1, \\ F^*(t), & \frac{1}{k+1} \le t < \frac{1}{k}, \quad k \in \mathbb{N} \\ 1, & t \ge 1, \end{cases} \tag{9.12}$$

where

$$F^*(t) = \frac{1}{1-p_1} \left\{ \sum_{s=1}^{k} (-1)^s \sum_{n=k+1}^{\infty} \binom{n}{s} (1-st)^{n-1} p_n + \mathbb{P}(N \ge k+1) \right\}.$$

By taking the derivative, we obtain the density of the continuous part,

$$f(t) = \frac{1}{1-p_1} \sum_{s=1}^{k} s(-1)^{s+1} \sum_{n=k+1}^{\infty} \binom{n}{s} (n-1)s(1-st)^{n-2} p_n, \tag{9.13}$$

for $\frac{1}{k+1} \le t < \frac{1}{k}$, $k \in \mathbb{N}$. The following result summarizes this discussion.

Theorem 9.1. *The CDF of R defined by (9.2) is given by*

$$G(t) = (1-p_1)F(t) + p_1 \mathbf{1}_{[1,\infty)}(t), \quad t \in \mathbb{R}, \tag{9.14}$$

with F given by (9.12), so that

$$R \stackrel{d}{=} I + (1-I)X, \tag{9.15}$$

where X has the CDF (9.12) and I is an indicator variable with $\mathbb{P}(I = 1) = p_1$, independent of X.

9.3.2 Peak to Average Ratio

We now turn to a derivation of the PTARN distribution of (9.3). Proceeding again with a standard conditioning argument, we obtain

$$\mathbb{P}(\tilde{R} \le x) = \sum_{n=1}^{\infty} \mathbb{P}(\tilde{R} \le x | N = n) \mathbb{P}(N = n) = \sum_{n=1}^{\infty} H_n(x/n) p_n,$$

with H_n and p_n given by (9.7) and (9.10), respectively. Since $H_n(u)$ is zero for $u < 1/n$, the probability $\mathbb{P}(\tilde{R} \le x)$ is zero whenever $x < 1$. Assume that $x \ge 1$, with $[x] = m \in \mathbb{N}$, so that $m \le x < m + 1$. Since $H_n(u) = 1$ whenever $u \ge 1$, we have $H_n(x/n) = 1$ whenever $n \le m = [x]$. On the other hand, for $n > m$, we have $1/(k+1) < x/n \le 1/k$ with $k = [n/x]$, so that in view of (9.7) we obtain

$$H_n(x/n) = \sum_{s=1}^{[n/x]} \binom{n}{s} (-1)^s (1 - sx/n)^{n-1} + 1.$$

This leads to the following expression for the CDF of $\tilde{R}$:

$$G(x) = \begin{cases} 0, & x < 1, \\ 1 + \sum_{n=[x]+1}^{\infty} p_n \sum_{s=1}^{[n/x]} \binom{n}{s} (-1)^s (1 - sx/n)^{n-1} & x \ge 1. \end{cases} \tag{9.16}$$

Note that $G(t)$ is continuous except for a jump of p_1 at $x = 1$, since $\tilde{R} = 1$ whenever $N = 1$, which occurs with probability p_1. Thus, the distribution of $\tilde{R}$ is a mixture of a point mass at 1 and a continuous distribution on $(1, \infty)$, the latter given by the CDF

$$F(x) = \begin{cases} 0, & x < 1, \\ \frac{1}{1-p_1} \left\{ 1 - p_1 + \sum_{n=[x]+1}^{\infty} p_n \sum_{s=1}^{[n/x]} \binom{n}{s} (-1)^s (1 - sx/n)^{n-1} \right\} & x \ge 1. \end{cases} \tag{9.17}$$

We summarize this discussion in the following result.

Theorem 9.2. *The CDF of $\tilde{R}$ defined by (9.3) is given by*

$$G(t) = (1 - p_1) F(t) + p_1 \mathbf{1}_{[1,\infty)}(t), \quad t \in \mathbb{R}, \tag{9.18}$$

with F given by (9.17), so that

$$\tilde{R} \stackrel{d}{=} I + (1 - I) X, \tag{9.19}$$

where X has the CDF (9.17) and I is an indicator variable with $\mathbb{P}(I = 1) = p_1$, independent of X .

9.4 Geometric Example

In this section, we illustrate our general results with an example of geometrically distributed N with PDF given by (9.4). Let us start with the PTSRN distribution, denoted by $PTSRN(p)$. By replacing the quantity p_n in (9.12) with the geometric PDF (9.4), we obtain the following expression for the CDF of the random variable

X of Theorem 9.1:

$$F(t) = \frac{p}{1-p}\sum_{s=1}^{k}(-1)^s \sum_{n=k+1}^{\infty}\binom{n}{s}(1-q)^{n-1}+(1-p)^{k-1}, \quad \frac{1}{k+1} \le t < \frac{1}{k}, k\in\mathbb{N}, \tag{9.20}$$

where

$$q = q_s(t;p) = 1-(1-p)(1-ts) \in (0,1). \tag{9.21}$$

The following result provides an explicit expression for the infinite sum in (9.20).

Lemma 9.1. *For any $q \in (0,1)$, $k \in \mathbb{N}$, and $s \in \{1,2,\ldots,k\}$ we have*

$$\sum_{n=k+1}^{\infty}\binom{n}{s}(1-q)^{n-1} = \frac{(1-q)^{s-1}}{q^{s+1}} - \sum_{n=s+1}^{k+1}\binom{n-1}{s}(1-q)^{n-2}. \tag{9.22}$$

This leads to the result below.

Theorem 9.3. *Let $R \sim PTSRN(p)$. Then the CDF of R is given by*

$$G(t) = (1-p)F(t) + p\mathbf{1}_{[1,\infty)}(t), \quad t \in \mathbb{R}, \tag{9.23}$$

where $F(t) = 0$ for $t \le 0$, $F(t) = 1$ for $t \ge 1$, while for $\frac{1}{k+1} \le t < \frac{1}{k}, k \in \mathbb{N}$,

$$F(t) = \frac{p}{1-p}\sum_{s=1}^{k}(-1)^s\left[\frac{(1-q)^{s-1}}{q^{s+1}} - \sum_{n=s+1}^{k+1}\binom{n-1}{s}(1-q)^{n-2}\right]+(1-p)^{k-1}, \tag{9.24}$$

with $q = q_s(x;p)$ given by (9.21).

Remark 9.4. As before, the random variable R admits the mixture representation (9.15), where the continuous variable X has the CDF F above. Moreover, by taking the derivative, we find the PDF of X to be

$$f(t) = p\sum_{s=1}^{k}s(-1)^{s+1}\left[\frac{(s+1-2q)(1-q)^{s-2}}{q^{s+2}} - \sum_{n=s+1}^{k+1}\binom{n-1}{s}(n-2)(1-q)^{n-3}\right] \tag{9.25}$$

for $\frac{1}{k+1} \le t < \frac{1}{k}, k \in \mathbb{N}$, and zero otherwise.

Finally, we consider the peak-to-average ratio (9.3) with geometrically distributed N, which distribution shall be denoted by $PTARN(p)$. When we replace the quantity p_n in (9.17) with the geometric PDF (9.4), we obtain the following expression for the CDF of the random variable X of Theorem 9.2:

$$F(x) = \begin{cases} 0, & x < 1, \\ 1 + \frac{p}{1-p} \sum_{n=[x]+1}^{\infty} \sum_{s=1}^{[n/x]} \binom{n}{s} (-1)^s [(1-p)(1-sx/n)]^{n-1} & x \geq 1. \end{cases} \tag{9.26}$$

In this case the mixture representation (9.19) still holds with $\mathbb{P}(I = 1) = p$ and X having a continuous distribution on $(1, \infty)$ given by the CDF (9.26).

9.5 An Illustrative Data Example

We illustrate the possible applications of the models for the ratio of the maximum to the total of IID exponentials with an example from climatology. Understanding the distribution of the ratio of maximum to the total precipitation over a time period helps with water management and flood control (see e.g., [19, 22, 23], or [24] and the references therein). The issue of flooding is especially important to study in the "low countries", such as the Netherlands. The latter is a country with about 25% of its area below the mean sea level, making 56% of its population living on the floodplains (see [25]) and 70% of the country gross domestic product produced in these low areas. Through the Netherlands' history, countless people have lost their lives and homes to floods from the North Sea or the rivers that could not be held by the flood defenses. The most recent flood events in the Netherlands were the 1993 and 1995 floods and the disastrous North Sea Flood of 1953 when 1,836 people were killed (see [3]).

For this illustration, we chose precipitation data from two stations nearby a massive river delta in the southwestern Netherlands, west from Antwerp: station 745 in Axel and station 770 in Westdorpe. Axel has data records from 1906–1995 and Westdorpe from 1995–2008. The two towns are about 4 miles apart. The first station is located at 51°17′ latitude, 3°55′ longitude, and 2.1 (m) elevation, while the Westdorpe station is located at 51°13′ latitude, 3°51′ longitude, and 1.5 (m) elevation. The precipitation is measured at 08.00 UT, as a 24 h total (from 08.00 to 08.00 UT), see "`http://www.knmi.nl/`" for more details. The precipitation measurement unit is 0.1 mm. For modeling, the data consisted of 4,197 weekly precipitation totals (from 01-01-1928 to 06-07-2008). Each four consecutive weeks were treated as 1 month. One month with zero total precipitation was excluded from analysis. We had 1,049 months in the data set . For each month two variables were computed: monthly total precipitation and monthly maximum precipitation. Monthly total was the sum of the four weekly totals and monthly maximum was the largest weekly total for the 4 weeks in the month. The data were then converted to the ratios of the maximum to the total, resulting in 1,049 values.

We fit the distribution PTSRn with $n = 4$ to the precipitation ratio data. The theoretical PDF and CDF are given by (9.5) and (9.7), respectively, with $n = 4$. Figure 9.2 presents graphical goodness-of-fit analysis: a histogram of the data overlayed with the theoretical PDF, a comparison of theoretical and empirical CDFs, and a

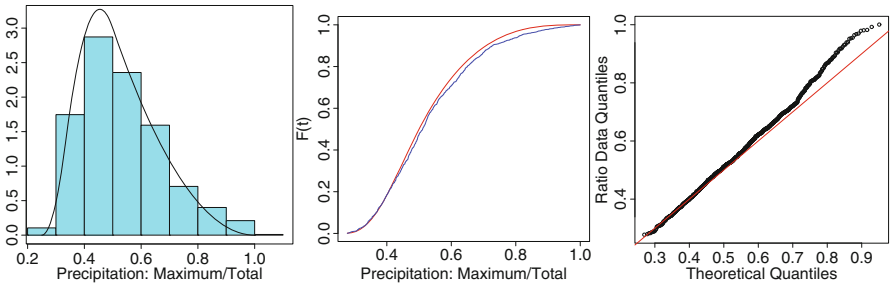


Fig. 9.2 Ratio of maximum to total monthly precipitation goodness-of-fit plots: histogram with fitted PDF (*left panel*), empirical versus theoretical CDF (*middle panel*), and QQ-plot (*right panel*)

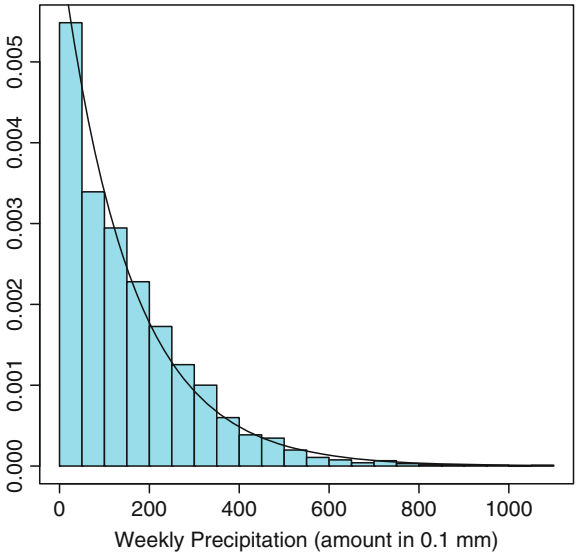


Fig. 9.3 Weekly precipitation amount data overlayed with a fitted exponential PDF

Quantile-Quantile (QQ) plot. The graphs show that the fit of the theoretical model to the data is quite reasonable.

We also examined the assumptions of the model by testing whether the observations (weekly total precipitation amounts) are distributed exponentially, using the likelihood ratio test discussed in [14, 19]. The null hypothesis is that the data comes from an exponential distribution, versus the alternative that the data comes from a Pareto distribution. Results from the likelihood ratio test indicated that the weekly precipitation amounts can be assumed to follow an exponential distribution (p-value $= 0.50$), with an estimated parameter of $\hat{\beta} = 0.00648$. Figure 9.3 provides histogram of the weekly precipitation amounts overlayed with the exponential density curve with $\beta = 0.00648$. The graphical fit confirms the results of the goodness-of-fit test as the density curve closely follows the histogram.

Additionally, we compared the empirical and theoretical correlation coefficients between the monthly maximum and the sum. The empirical correlation coefficient between the sum and maximum precipitation was found to be $\hat{r} = 0.871$. According to the results of [20], the theoretical correlation between them should be $r = 0.873$, which is rather close to the empirical one.

Our example illustrates that the models involving the ratio of the maximum and the sum of exponential variables are not only an important theoretically, but are also promising as stochastic models in actual applications.

Appendix

Proof of Lemma 9.1. Write the left-hand-side of (9.22) as

$$\frac{(1-q)^{s-1}}{q^{s+1}} \sum_{j=k+2}^{\infty} \binom{j-1}{s+1-1} q^{s+1}(1-q)^{j-(s+1)}. \tag{9.27}$$

Note that this is the same as the probability $\mathbb{P}(T \geq k+2)$, where T is a negative binomial random variable given by the PDF

$$\mathbb{P}(T = j) = \binom{j-1}{s+1-1} q^{s+1}(1-q)^{j-(s+1)}, \quad j = s+1, s+2, \ldots.$$

Since $\mathbb{P}(T \geq k+2) = 1 - \mathbb{P}(T \leq k+1)$, we have

$$\sum_{j=k+2}^{\infty} \binom{j-1}{s+1-1} q^{s+1}(1-q)^{j-(s+1)} = 1 - \sum_{n=s+1}^{k+1} \binom{n-1}{s+1-1} q^{s+1}(1-q)^{n-(s+1)},$$

and the result follows by substituting the right-hand-side above into (9.27). □

Acknowledgements This work was supported in part by NSF grants ATM-0236898 and ATM-0503722.

References

1. Aebi M, Embrechts P, Mikosch T (1992) A large claim index. Bull Swiss Assoc Actuar 2: 143–156
2. Arov DZ, Bobrov AA (1960) The extreme terms of a sample and their role in the sum of independent random variables. Theory Probab Appl 5:377–396
3. Asselman NEM, Jonkman SN (eds) (2007) A method to estimate loss of life caused by large-scale floods in the Netherlands. In: Flood risk management in Europe Advances in natural and technological hazards research, vol 25. Springer, The Netherlands, pp 155–170

4. Breiman L (1965) On some limit theorems similar to the arc-sine law. Theory Probab Appl 10:323–331
5. Darling DA (1952) The influence of the maximum term in the sum of independent random variables. Trans A Math Soc 73:95–107
6. Downey PJ, Wright PE (1994) The ratio of the extreme to the sum in a random sequence with applications. Technical Report 94-18, University of Arizona, Tucson, AZ, 85721
7. Fieller NRJ (1976) PhD thesis, Department of Mathematics and Statistics, University of Hull
8. Fisher RA (1929) Tests of significance in harmonic analysis. Proc R Soc Lond A 125:54–59
9. Haas PJ (1992) The maximum and mean of a random length sequence. J Appl Probab 29: 460–466
10. Kesten H, Maller RA (1994) Infinite limits and infinite limit points of random walks and trimmed sums. Ann Probab 22(93):1473–1513
11. Kimber AC (1979) Test for a single outlier in a gamma sample with unknown shape and scale parameters. Appl Stat 28:243–250
12. Kimber AC (1982) Tests for many outliers in an exponential sample. Appl Stat 31(3):263–271
13. Kimber AC, Stevens HJ (1981) The null distribution of a test for two upper outliers in an exponential sample. Appl Stat 30(2):153–157
14. Kozubowski TJ, Panorska AK, Qeadan F, Gershunov A, Rominger D (2009) Testing exponentiality versus Pareto distribution via likelihood ratio. Commun Stat Simul Comput 38(1):118–139
15. Lewis T, Fieller NRJ (1979) A recursive algorithm for null distributions for outliers: I. Gamma samples. Technometrics 21:371–376
16. Maller RA, Resnick SI (1984) Limiting behaviour of sums and the term of maximum modulus. Proc Lond Math Soc (3) 49:385–422
17. Morrison M, Tobias F (1965) Some statistical characteristics of peak to average ratio. Technometrics 7(3):379–385
18. O'Brien GL (1980) A limit theorem for sample maxima and heavy branches in Galton-Watson trees. J Appl Probab 17:281–291
19. Panorska AK, Gershunov A, Kozubowski TJ (2007) From diversity to volatility: Probability of daily precipitation extremes. In: Tsonis A, Elsner J (eds) Nonlinear dynamics in geosciences. Springer, New York, pp 465–484
20. Qeadan F, Kozubowski TJ, Panorska AK (2008) The joint distribution of the sum and the maximum of n i.i.d. exponential random variables, preprint
21. Rajbanshi R, Wyglinski A, Minden G (2006) Subcarrier power adjustment technique for peak-to-average power ratio reduction of OFDM systems. Proceedings of the IEEE military communications conference, Washington, DC, USA
22. Singh VP (ed) (1995) Environmental hydrology. Water science and technology library, Springer, Berlin
23. Unkasevic M, Radinovic D (2000) Statistical analysis of daily maximum and monthly precipitation at Belgrade. Theor Appl Climatol 66:241–249
24. VanRheenen NT, Palmer RN, Hahn MA (2003) Evaluating potential climate change impacts on water resources systems operations: Case studies of Portland, Oregon and Central Valley, California. Water Resour Update 124:35–50
25. Ven van de GP (1993) Man-made lowlands: History of water management and land reclamation in the Netherlands. Uitgeverij Matrijs: Utrecht, The Netherlands

Chapter 10
Least Square Estimation for Regression Parameters Under Lost Association

Vasudevan Mangalam

Abstract A method is developed to deal with estimating the regression coefficients when the association among the paired data is partially or completely lost. Asymptotic properties of the estimators are discussed.

10.1 Introduction

Consider a situation where we have two data sets that are believed to be linearly related. The independent variable is denoted by x and the dependent variable is denoted by y. The classical regression analysis gives us least square estimators for the slope and intercept parameters that are consistent and asymptotically normal. In this paper the author looks at a situation where the complete data sets for the independent and dependent variables are available, but the pairwise association between them is lost. In other words, we have a set of x observations and a set of y observations without knowing which value of the dependent variable was originally associated with which value of the independent variable. Such a situation can arise out of accidental sorting of values that destroyed the association or when the two data were collected separately or by different people.

The model we work with is the standard linear regression model. Let X_1, $X_2, \ldots, X_n$ be i.i.d. with mean μ_X and standard deviation σ_X. Let $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ be i.i.d. with mean zero and standard deviation σ_ϵ so that ϵ_i's are independent of X_i's. Finally, let $Y_i = \alpha + \beta X_i + \epsilon_i$, where α and β are unknown parameters. The observed data consists of $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_n\}$ without any information on the pairing. In absence of association, it seems natural that we should define the least square estimator by choosing the permutation of y_i's that would minimize the error sum of squares. If $\Phi(\alpha, \beta) = \min_{\pi \in P_n} \sum_{i=1}^{n} (y_{\pi(i)} - (\alpha + \beta x_i))^2$ where P_n represent the set of all permutations of $\{1, 2, \ldots, n\}$, then our estimators

V. Mangalam
Department of Mathematics, Universiti Brunei Darussalam, Gadong BE 1410, Brunei Darussalam
e-mail: vmangalam@gmail.com

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_10,

of α and β are $\hat{\alpha}$ and $\hat{\beta}$ that satisfy the equation

$$\Phi(\hat{\alpha}, \hat{\beta}) = \min_{\alpha,\beta} \Phi(\alpha, \beta).$$

10.2 Main Results

In what follows, for any variable x, $x_{n:i}$ represents the ith order statistics of $\{x_1, x_2, \ldots, x_n\}$.

Lemma 10.1.

$$\Phi(\hat{\alpha}, \hat{\beta}) = \min_{\pi \in P_n} S_{\mathbf{yy}}(1 - r^2_{\mathbf{x}\pi(\mathbf{y})}) \tag{10.1}$$

where $S_{\mathbf{yy}} = \sum_{i=1}^{n} (y_i - \bar{y})^2$, $r_{\mathbf{xy}}$ *indicates the correlation between* $\mathbf{x}$ *and* $\mathbf{y}$, *and* $\pi(\mathbf{y})$ *represents the vector obtained by rearranging the values of the vector* $\mathbf{y}$ *according to the permutation* π.

Proof. Let A represent the RHS of (10.1) and let the minimum in (10.1) be attained by the permutation π_0. By the results of standard regression analysis, (see, for instance, [2]) we know that for any pair of vectors $\mathbf{x}$ and $\mathbf{y}$, $\min_{\alpha,\beta} \sum_{i=1}^{n} \left[y_i - (\alpha + \beta x_i)\right]^2 = S_{\mathbf{yy}}(1 - r^2_{\mathbf{xy}})$. So for any permutation π, $\min_{\alpha,\beta} \sum_{i=1}^{n} \left[y_{\pi(i)} - (\alpha + \beta x_i)\right]^2 = S_{\pi(\mathbf{y})\pi(\mathbf{y})}(1 - r^2_{\mathbf{x}\pi(\mathbf{y})}) = S_{\mathbf{yy}}(1 - r^2_{\mathbf{x}\pi(\mathbf{y})})$.

$$\begin{aligned}
A &= S_{\mathbf{yy}}(1 - r^2_{\mathbf{x}\pi_0(\mathbf{y})}) \\
&= \min_{\alpha,\beta} \sum_{i=1}^{n} (y_{\pi_0(i)} - (\alpha + \beta x_i))^2 \\
&\geq \min_{\alpha,\beta} \min_{\pi \in P_n} \sum_{i=1}^{n} (y_{\pi(i)} - (\alpha + \beta x_i))^2 \\
&= \min_{\alpha,\beta} \Phi(\alpha, \beta) \\
&= \Phi(\hat{\alpha}, \hat{\beta}).
\end{aligned}$$

For any $\pi \in P_n$,

$$\begin{aligned}
A &\leq S_{\mathbf{yy}}(1 - r^2_{\mathbf{x}\pi(\mathbf{y})}) \\
&= \min_{\alpha,\beta} \sum_{i=1}^{n} (y_{\pi(i)} - (\alpha + \beta x_i))^2.
\end{aligned}$$

Thus $A \leq \sum_{i=1}^{n} (y_{\pi(i)} - (\alpha + \beta x_i))^2$ $\forall \alpha$ and β and $\forall \pi \in P_n$. Therefore $\forall \alpha, \beta$,

$$\begin{aligned} A &\leq \min_{\pi \in P_n} \sum_{i=1}^{n} (y_{\pi(i)} - (\alpha + \beta x_i))^2 \\ &= \Phi(\alpha, \beta). \end{aligned}$$

So $A \leq \min_{\alpha,\beta} \Phi(\alpha, \beta) = \Phi(\hat{\alpha}, \hat{\beta})$. □

One of the consequences of Lemma 10.1 is that the permutation that minimizes Φ is not dependent on the values of α and β. If we find the permutation π that maximizes $r^2_{\mathbf{x}\pi(\mathbf{y})}$, then it minimizes $S_{\mathbf{yy}}(1 - r^2_{\mathbf{x}\pi(\mathbf{y})})$ and consequently, $\hat{\beta} = S_{\mathbf{x}\pi(\mathbf{y})}/S_{\mathbf{xx}}$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

The difficulty is, of course, the computational complexity. If we were to try out all $n!$ permutations the computational time required explodes at a faster-than-exponential rate. This problem is tackled by Theorem 10.1. Before that, we need a preliminary lemma.

Lemma 10.2. *Let $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ be two sets of real numbers such that $x_1 \leq x_2 \leq \ldots, \leq x_n$, $y_1 \leq y_2 \leq \ldots, \leq y_n$. Let P_n represent the set of all permutations of $1, 2, \ldots, n$. Then*

(1) $\max_{\pi \in P_n} \sum_{i=1}^{n} x_i y_{\pi(i)} = \sum_{i=1}^{n} x_i y_i$

(2) $\min_{\pi \in P_n} \sum_{i=1}^{n} x_i y_{\pi(i)} = \sum_{i=1}^{n} x_i y_{n-i+1}$.

Proof. (1) The proof is by induction. We first prove the result for $n = 2$.

$$\begin{aligned} 0 &\leq (x_2 - x_1)(y_2 - y_1) = x_1 y_1 + x_2 y_2 - (x_1 y_2 + x_2 y_1) \\ &\Rightarrow x_1 y_2 + x_2 y_1 \leq x_1 y_1 + x_2 y_2. \end{aligned}$$

Now assume the result to be true for integers up to and including $n - 1$. Letting π to be an arbitrary element in P_n, we will prove that

$$\sum_{i=1}^{n} x_i y_{\pi(i)} \leq \sum_{i=1}^{n} x_i y_i.$$

Let $j = \pi(n)$ and $k = \pi^{-1}(n)$ so that $\pi(k) = n$. Then

$$\begin{aligned} \sum_{i=1}^{n} x_i y_{\pi(i)} &= \sum_{\substack{i=1 \\ i \neq k}}^{n-1} x_i y_{\pi(i)} + x_k y_n + x_n y_j \\ &= \left\{ \sum_{\substack{i=1 \\ i \neq k}}^{n-1} x_i y_{\pi(i)} + x_k y_j \right\} + x_k y_n + x_n y_j - x_k y_j. \end{aligned}$$

The expression in the braces, say B, is of the form $\sum_{i=1}^{n-1} x_i y_{\pi(i)}$ and hence by induction hypothesis, $B \leq \sum_{i=1}^{n-1} x_i y_i$. Also, applying the result for $n = 2$ with x_k, x_n, y_j and y_n, we get $x_n y_j + x_k y_n \leq x_k y_j + x_n y_n$. Hence

$$\begin{aligned}\sum_{i=1}^{n} x_i y_{\pi(i)} &\leq \sum_{i=1}^{n-1} x_i y_i + x_k y_n + x_n y_j - x_k y_j \\ &= \sum_{i=1}^{n-1} x_i y_i + x_n y_n \\ &= \sum_{i=1}^{n} x_i y_i.\end{aligned}$$

(2) Apply Part (1) with y_i's replaced by their negatives. □

Define two permutations R_X and $\bar{R}_X$ as follows. $R_X(i)$ is the rank of x_i among $\{x_1, x_2, \ldots\}$ and $\bar{R}_X(i) = n - R_X(i) + 1$. Let R_Y and $\bar{R}_Y$ be defined similarly. Now define permutations π_1 and π_2 by $\pi_1 = R_Y^{-1} R_X$ and $\pi_2 = \bar{R}_Y^{-1} R_X$.

Theorem 10.1. *If π is such that $r^2_{\mathbf{x}\pi(\mathbf{y})} = \max\{r^2_{\mathbf{x}\pi_1(\mathbf{y})}, r^2_{\mathbf{x}\pi_2(\mathbf{y})}\}$, then $\hat{\beta} = S_{x\pi(y)}/S_{xx}$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.*

Proof. We shall show that

$$\Phi(\hat{\alpha}, \hat{\beta}) = S_{\mathbf{yy}}(1 - \max\{r^2_{\mathbf{x}\pi_1(\mathbf{y})}, r^2_{\mathbf{x}\pi_2(\mathbf{y})}\}),$$

from which the theorem follows.

From Lemma 10.1, $\Phi(\hat{\alpha}, \hat{\beta}) = \min_{\pi \in P_n} S_{\mathbf{yy}}(1 - r^2_{\mathbf{x}\pi(\mathbf{y})}) = S_{\mathbf{yy}}(1 - \max_{\pi \in P_n} r^2_{\mathbf{x}\pi(\mathbf{y})})$, so it is sufficient to prove that $\max_{\pi \in P_n} r^2_{\mathbf{x}\pi(\mathbf{y})} = \max\{r^2_{\mathbf{x}\pi_1(\mathbf{y})}, r^2_{\mathbf{x}\pi_2(\mathbf{y})}\}$. Moreover, $\max_{\pi \in P_n} r^2_{\mathbf{x}\pi(\mathbf{y})} = \max\{(\max_{\pi \in P_n} r_{\mathbf{x}\pi(\mathbf{y})})^2, (\min_{\pi \in P_n} r_{\mathbf{x}\pi(\mathbf{y})})^2\}$.

Note that $\sum_{i=1}^{n} x_i y_{\pi_1(i)} = \sum_{i=1}^{n} x_{n:i} y_{n:i}$. Now from Lemma 10.2,

$$\begin{aligned}\max_{\pi \in P_n} r_{\mathbf{x}\pi(\mathbf{y})} &= \left[\max_{\pi \in P_n} \frac{1}{n} \sum_{i=1}^{n} (x_i y_{\pi(i)} - \bar{x}\bar{y})\right] \Big/ s_x s_y \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} x_{n:i} y_{n:i} - \bar{x}\bar{y}\right) \Big/ s_x s_y \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_{\pi_1(i)} - \bar{x}\bar{y}\right) \Big/ s_x s_y \\ &= r_{\mathbf{x}\pi_1(\mathbf{y})}.\end{aligned}$$

Similarly, $\min_{\pi \in P_n} r_{\mathbf{x}\pi(\mathbf{y})} = r_{\mathbf{x}\pi_2(\mathbf{y})}$. Thus $\max_{\pi \in P_n} r^2_{\mathbf{x}\pi(\mathbf{y})} = \max\{r^2_{\mathbf{x}\pi_1(\mathbf{y})}, r^2_{\mathbf{x}\pi_2(\mathbf{y})}\}$. □

What we infer from the preceding results is that the least-square estimators are obtained by ordering the x-values in the ascending order, ordering y-values in ascending order or descending order depending on which gives a higher coefficient of determination and finding the standard estimators of the parameters based on this arrangement. But note that the sign of the slope parameter can be wrongly chosen. It is possible that a strongly positively correlated bivariate data set yields a higher coefficient of determination when the variables are ordered in reverse and hence the method can lead to wrong conclusions on whether β is positive or negative. Thus it would be prudent to use prior information on the sign of β.

10.3 Almost Sure Limits

In order to find the limits of least square estimators of the regression parameters, we need the following theorem from [1]. The details of the assumptions are described in the Appendix.

Theorem 10.2. *Let* $\left\{(X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(d)}),\ i = 1, 2, \ldots\right\}$ *be a sequence of random vectors such that for each* j $(1 \le j \le d)$, $\left\{X_1^{(j)}, X_2^{(j)}, \ldots, \right\}$ *forms a sequence of i.i.d. random variables with continuous distribution function* F_j. *Let* ϕ *be a real-valued measurable function on* $\mathbb{R}^d$ *such that* ϕ *satisfies some regularity conditions. Then*

$$\frac{1}{n}\sum_{i=1}^{n} \phi\left(X_{n:i}^{(1)}, \ldots, X_{n:i}^{(d)}\right) \xrightarrow{a.s.} \gamma \equiv E\phi\left(F_1^{-1}(U), F_2^{-1}(U), \ldots, F_d^{-1}(U)\right)$$

as $n \to \infty$, *where* U *is uniformly distributed over* $(0, 1)$.

Note that in the above theorem, no assumption is made about the joint distribution of X_i and Y_i. We are now ready to prove the following lemma.

Lemma 10.3. *Let* F *be a continuous distribution function with mean* μ *and standard deviation* σ. *Let* $X_1,\ X_2,\ \ldots,\ X_n$ *be i.i.d. with distribution function* F, *and* $Y_1,\ Y_2,\ \ldots,\ Y_n$ *be i.i.d. with distribution function* F. *Then,*

$$\frac{1}{n}\sum_{i=1}^{n} X_{n:i} Y_{n:i} \to^{a.s.} \mu^2 + \sigma^2$$

and

$$\frac{1}{n}\sum_{i=1}^{n} X_{n:i} Y_{n:n-i+1} \to^{a.s.} \mu^2 - \sigma^2.$$

Proof. Apply Theorem 10.2 with $n = 2$ and $\phi(x, y) = xy$, which satisfies the regularity conditions. Here the limit γ is given by

$$\begin{aligned}\gamma &= E\phi\left(F^{-1}(U), F^{-1}(U)\right)\\ &= \int_0^1 (F^{-1}(u))^2 du\\ &= \int_{-\infty}^{\infty} x^2 dF(x)\\ &= \mu^2 + \sigma^2.\end{aligned}$$

The second part of the lemma follows by applying Theorem 10.2 to X and $-Y$. □

Lemma 10.4. *Let F and G be continuous distribution functions such that there exists a, b such that $G(x) = F(\frac{x-a}{b})$. Assume F has mean μ_1 and standard deviation σ_1, and G has mean μ_2 and standard deviation σ_2. Let $X_1, X_2, \ldots, X_n$ be i.i.d. with distribution function F and $Y_1, Y_2, \ldots, Y_n$ be i.i.d. with distribution function G. Then,*

$$\frac{1}{n}\sum_{i=1}^{n} X_{n:i} Y_{n:i} \to^{a.s.} \mu_1\mu_2 + \sigma_1\sigma_2$$

$$\frac{1}{n}\sum_{i=1}^{n} X_{n:i} Y_{n:n-i+1} \to^{a.s.} \mu_1\mu_2 - \sigma_1\sigma_2.$$

Proof. The result follows from Lemma 10.3 by a linear transformation. □

Next, we look at the asymptotic properties of the estimator. We would hope that the estimator is consistent and asymptotically normal. Unfortunately, when the association between x and y is completely lost, we do not get consistency or unbiasedness, nor are we able to recover the sign of β. The limiting behaviour of the least square estimators are given in the following theorem.

Theorem 10.3.

$$\hat{\beta}_n \to^{a.s.} \beta_1 \equiv \pm\sqrt{\beta^2 + \frac{\sigma_\epsilon^2}{\sigma_x^2}}.$$

$$\hat{\alpha}_n \to^{a.s.} \alpha_1 \equiv \alpha + \mu_x(\beta - \beta_1).$$

Proof. From Lemma 10.4,

$$\begin{aligned}s_{x\pi_1(y)} &= \frac{1}{n}\sum_{i=1}^{n} x_{n:i} y_{n:i} - \bar{x}\bar{y}\\ &\to^{a.s.} E(X)E(Y) + SD(X)SD(Y) - E(X)E(Y)\\ &= SD(X)SD(Y).\end{aligned}$$

Similarly $s_{x\pi_2(y)} \to^{a.s.} -SD(X)SD(Y)$. As $s_{xn}^2 \to^{a.s.} V(X)$ and $\hat{\beta}_n = s_{x\pi(y)}/s_{xn}^2$ where π is π_1 or π_2,

$$\begin{aligned}\hat{\beta}_n \to^{a.s.} & \pm \frac{SD(Y)}{SD(X)} = \pm\sqrt{\beta^2 + \frac{\sigma_\epsilon^2}{\sigma_x^2}} \\ \hat{\alpha}_n &= \bar{y}_n - \hat{\beta}_n \bar{x}_n \\ &\to^{a.s.} \mu_y - \beta_1 \mu_x \\ &= \alpha + \mu_x (\beta - \beta_1).\end{aligned}$$

□

10.4 Asymptotic Normality

For the asymptotic normality to hold for the least square estimators, we need to make certain assumptions about the distributions of X and ϵ. In order to make matters simple, we will not deal here with the case where X and ϵ have general continuous distributions. We shall assume that both are normally distributed, as a consequence of which (X, Y) has bivariate normal distribution.

We need to use another theorem from [1] as stated below. Once again, details are given in the Appendix.

Theorem 10.4. *Let $\mathbf{X}_i = (X_i^{(1)}, \ldots, X_i^{(d)})$ be i.i.d. random vectors. Let F_j $(1 \le j \le d)$ denote the marginal distribution of $X_i^{(j)}$ which is assumed to be continuous, and $F_{j,k}$ $(1 \le j, k \le d)$ denote the marginal distribution of $(X_i^{(j)}, X_i^{(k)})$. If ϕ satisfies some regularity conditions, then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi\left(X_{n:i}^{(1)}, \ldots, X_{n:i}^{(d)}\right) - \sqrt{n}\ \gamma \to^d N(0, \sigma^2) \tag{10.2}$$

as $n \to \infty$ where γ is as in Theorem 10.2 and

$$\begin{aligned}\sigma^2 = 2 \sum_{j=1}^{d} \int_0^1 \int_0^y x(1-y)\psi_j(x)\psi_j(y)dxdy \\ + 2 \sum_{1 \le j < k \le d} \int_0^1 \int_0^1 [G_{j,k}(x, y) - xy]\psi_j(x)\psi_k(y)dxdy\end{aligned}$$

where $G_{j,k}(x, y) = F_{j,k}\left(F_j^{-1}(x), F_k^{-1}(y)\right)$.

Lemma 10.5. *Let (X_1, Y_1), (X_2, Y_2), $\ldots$, (X_n, Y_n) be i.i.d. bivariate normal with each component being $N(0, 1)$. Then,*

1. $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_{n:i} - Y_{n:i})^2 \to^p 0$
2. $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[X_{n:i} Y_{n:i} - \overline{X}_n \overline{Y}_n - 1\right] \to^d N\left(0, \tau^2\right)$
   where $\tau^2 = \frac{1}{2}\left[E\left(X^2 Y^2\right) + 1\right]$.

Proof. For Part 1, apply Theorem 10.4 with $n = 2$, $\phi(x, y) = (x - y)^2$. It can be verified that when the distributions are normal, the conditions of Theorem 10.2 and Theorem 10.4 are satisfied. It is easily seen that for this case, the asymptotic mean and variance are both zero, and hence the result holds. As for Part 2,

$$2X_{n:i}Y_{n:i} = X_{n:i}^2 + Y_{n:i}^2 - (X_{n:i} - Y_{n:i})^2,$$

so

$$2\sum_{i=1}^{n} X_{n:i}Y_{n:i} = \sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} (X_{n:i} - Y_{n:i})^2$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[X_{n:i}Y_{n:i} - \overline{X}_n\overline{Y}_n - 1\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{X_i^2 + Y_i^2}{2} - 1\right] + \sqrt{n}\,\overline{X}_n\overline{Y}_n + o_p(1)$$

from Part 1. By the Lindeberg–Lévy central limit theorem (See [3], for instance), the first term of RHS goes to normal with mean zero and variance equaling

$$\begin{aligned} V\left(\frac{X^2 + Y^2}{2}\right) &= \frac{1}{4}\left[V\left(X^2\right) + V\left(Y^2\right) + 2COV\left(X^2, Y^2\right)\right] \\ &= \frac{1}{2}\left[2 + COV\left(X^2, Y^2\right)\right] \\ &= \frac{1}{2}\left[E\left(X^2Y^2\right) + 1\right] \\ &= \tau^2. \end{aligned}$$

$\sqrt{n}\,\overline{X}_n\overline{Y}_n \to^p 0$ because $\sqrt{n}\,\overline{X}_n$ is asymptotically normal, and hence $O_p(1)$, and $\overline{Y}_n \to^p 0$. The result now follows from Slutsky's theorem. □

Note: The asymptotic variance τ^2 is equal to $\frac{V(X^2)}{2}$ if X_i's are independent of Y_i's, whereas $\tau^2 = V\left(X^2\right)$ if $X_i = Y_i$ for all i.

Lemma 10.6. *Let* (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n) *be i.i.d. bivariate normal such that* X_1, X_2, …, X_n *are i.i.d.* $N(\mu_1, \sigma_1)$ *and* Y_1, Y_2, …, Y_n *be i.i.d.* $N(\mu_2, \sigma_2)$. *Then,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[X_{n:i}Y_{n:i} - \left(\overline{X}_n\overline{Y}_n + \sigma_1\sigma_2\right)\right] \to^d N\left(0, \eta^2\right)$$

where $\eta^2 = \frac{\sigma_1^2\sigma_2^2}{2}E\left[\left(\frac{X-\mu_1}{\sigma_1}\right)^2\left(\frac{Y-\mu_2}{\sigma_2}\right)^2 + 1\right]$.

Proof. Let $Z_i = \frac{X_i - \mu_1}{\sigma_1}$ and $W_i = \frac{Y_i - \mu_2}{\sigma_2}$. By Lemma 10.5,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[Z_{n:i}W_{n:i} - \overline{Z}_n\overline{W}_n - 1\right] \to^d N\left(0, \eta^2 t\right).$$

The left hand side is equal to $\frac{1}{\sqrt{n}\sigma_1\sigma_2}\sum_{i=1}^{n}\left[X_{n:i}Y_{n:i} - \left(\overline{X}_n\overline{Y}_n + \sigma_1\sigma_2\right)\right]$ and hence the weak convergence holds. The asymptotic variance is given by $\frac{\sigma_1^2\sigma_2^2}{2}\left[E\left(Z^2W^2\right) + 1\right]$ which equals η^2. □

We are now ready to prove the asymptotic normalities of $\hat{\beta}_n$ and $\hat{\alpha}_n$.

Theorem 10.5. *If (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) are i.i.d. bivariate normal, then the least square estimators of the regression parameters under lost association are asymptotically normal.*

1. $\sqrt{n}\left(\hat{\beta}_n - \beta_1\right) \to^d N(0, \cdot)$
2. $\sqrt{n}\left(\hat{\alpha_n} - \alpha_1\right) \to^d N(0, \cdot)$
   where β_1 and α_1 are as in Theorem 10.3

Proof. $\hat{\beta}_n = s_{x\pi(y)}/s_{xn}^2$ where π is π_1 or π_2. We shall prove the case of $\pi = \pi_1$. The other case is similar. (It may fluctuate between the two permutations giving positive or negative values for $\hat{\beta}_n$. In that case the subsequences will converge to β_1 or $-\beta_1$). As $s_{xn}^2 \to^{a.s.} \sigma_x^2$,

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_n - \beta_1\right) &= \sqrt{n}\left(\frac{s_{x\pi_1(y)}}{s_{xn}^2} - \frac{\sigma_x\sigma_y}{\sigma_x^2}\right)\\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} x_{n:i}\,y_{n:i} - \bar{x}\bar{y}/s_{xn}^2\right)\\ &\approx \frac{1}{\sqrt{n}\sigma_x^2}\sum_{i=1}^{n}\left[X_{n:i}Y_{n:i} - \left(\overline{X}_n\overline{Y}_n + \sigma_1\sigma_2\right)\right].\end{aligned}$$

The asymptotic normality of $\hat{\beta}_n$ now follows from Lemma 10.6. The asymptotic normality of $\hat{\alpha}_n$ follows from that of $\hat{\beta}_n$ and the fact that $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$. □

10.5 Partial Loss of Association

It is conceivable that only for part of the data the association between x and y is lost. For the rest, we may have the full data set with the one to one correspondence intact. In such situations we expect consistency and asymptotic normality for the least square estimator to hold provided the proportion of observations for which association is lost tends to zero.

As this is a rather severe restriction, the author suggests alternative estimators for β and α as follows. Let $s_{\epsilon n}^2$ be the standard estimator of the error variance, s_{xn}^2 be the standard estimator of the variance of the x variable, and β_n^* be the standard estimator of β. These estimators are calculated using the part of the data for which association is not lost.

$$\tilde{\beta} = sgn(\beta_n^*)\sqrt{\hat{\beta_n}^2 - \frac{s_{\epsilon n}^2}{s_{xn}^2}}.$$

Assuming that the number of observations for which association is intact tends to infinity as the sample size tends to infinity, this new estimator is consistent. This is a straight forward consequence of the previous results. Its asymptotic normality is, at present, not as clear.

10.6 Conclusion

There is scope for further studies on this topic. Investigating MLE's is one possible extension. Another is to generalize the idea of lost association. In general, loss of association does not have to be a total hit or total miss. It is possible that for each x_i, we have a set of y_i's that x_i may be associated with, and this set may range from a singleton to the entire set of y_i's. On one extreme we have full association information and on the other we have none. Least square estimation and maximum likelihood estimation for such a general case can be a subject of future investigation.

Appendix

The content of this appendix is an excerpt from [1].

Let $\{(X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(d)}),\ i = 1, 2, \ldots\}$ be a sequence of random vectors such that for each j $(1 \leq j \leq d)$, $\{X_1^{(j)}, X_2^{(j)}, \ldots, \}$ forms a sequence of independent and identically distributed (*i.i.d.*) random variables with distribution function F_j. Let $X_{n:i}^{(j)}$ denote the ith order statistic of $\{X_1^{(j)}, X_2^{(j)}, \ldots X_n^{(j)}\}$. Define, for $0 < t < 1$, $F^{-1}(t) = \inf\{x : F(x) \geq t\}$ for a distribution function F.

Let ϕ be a real-valued measurable function on $\mathbb{R}^d$. Let $\mathbf{x} = (x_1, \ldots, x_d)$ for $0 < x, x_1, \ldots, x_d < 1$. Define

$$\psi(\mathbf{x}) = \phi\left(F_1^{-1}(x_1), \ldots, F_d^{-1}(x_d)\right), \tag{10.3}$$

$$\gamma(x) = \psi(x, x, \ldots, x), \tag{10.4}$$

$$\psi_j(x) = \frac{\partial \psi(\mathbf{x})}{\partial x_j}|_{(x,\ldots,x)}, \quad 1 \leq j \leq d \tag{10.5}$$

$$\psi_{j,k}(\mathbf{x}) = \frac{\partial^2 \psi(\mathbf{x})}{\partial x_j \partial x_k}, \quad 1 \leq j, k \leq d. \tag{10.6}$$

We introduce conditions on ϕ.

(C1) The function $\psi(\mathbf{u})$ is continuous at $\mathbf{u} = (u, \ldots, u), 0 < u < 1$. That is, ψ is continuous at each point on the diagonal of $(0, 1)^d$.
(C2) There exist K and $c_0 > 0$ such that for $(x_1, \ldots, x_d) \in (0, c_0)^d \cup (1-c_0, 1)^d$,

$$|\psi(x_1, \ldots, x_d)| \le K \left(1 + \sum_{j=1}^{d} |\gamma(x_j)|\right).$$

(C3) The functions, $[x(1-x)]^{3/2}[\psi_j(x)]^2$ $(1 \le j \le d)$, and $[x(1-x)]^{3/2}$ $|\psi_{j,k}(x, \ldots, x)|$ $(1 \le j, k \le d)$ are Riemann integrable over $(0, 1)$.
(C4) For all large m, there exist $K = K(m) \ge 1$ and $\delta > 0$ such that

$$\begin{aligned} &|\psi(\mathbf{y}) - \psi(\mathbf{x}) - \langle \mathbf{y} - \mathbf{x}, \nabla\psi(\mathbf{x})\rangle| \\ &\le K \sum_{j,k=1}^{d} |(y_j - x)(y_k - x)| \, [1 + |\psi_{j,k}(\mathbf{x})|] \end{aligned}$$

if $\mathbf{x} = (x, \ldots, x), \mathbf{y} = (y_1, \ldots, y_d) \in (0, 1)^d$, $||\mathbf{y} - \mathbf{x}||_{\ell_1} < \delta$, and for $1 \le j \le d$, $y_j(1 - y_j) > x(1 - x)/m$. Here $\|\mathbf{y}\|_{\ell_1} := |y_1| + \cdots + |y_d|$ denotes the ℓ_1-norm of $\mathbf{y}$ and $\nabla\psi(\mathbf{x})$ the gradient of ψ.

Theorem 10.6. *Let $\{(X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(d)}),\ i = 1, 2, \ldots\}$ be a sequence of random vectors such that for each j $(1 \le j \le d)$, $\{X_1^{(j)}, X_2^{(j)}, \ldots, \}$ forms a sequence of i.i.d. random variables with continuous distribution function F_j. Suppose that ϕ satisfies the conditions (C1)–(C2), and that γ, as defined in (10.4), is Riemann integrable, then*

$$\frac{1}{n} \sum_{i=1}^{n} \phi\left(X_{n:i}^{(1)}, \ldots, X_{n:i}^{(d)}\right) \xrightarrow{a.s.} \bar{\gamma}$$

as $n \to \infty$. Here $\bar{\gamma} = \int_0^1 \gamma(y)dy = E\phi\left(F_1^{-1}(U), F_2^{-1}(U), \ldots, F_d^{-1}(U)\right)$ and U is uniformly distributed over $(0, 1)$.

Theorem 10.7. *Let $\mathbf{X}_i = (X_i^{(1)}, \ldots, X_i^{(d)})$ be i.i.d. random vectors. Let F_j $(1 \le j \le d)$ denote the marginal distribution of $X_i^{(j)}$ which is assumed to be continuous, and $F_{j,k}$, $(1 \le j, k \le d)$ the marginal distribution of $(X_i^{(j)}, X_i^{(k)})$. If ϕ satisfies conditions (C1)–(C4), and that γ, as defined in (10.4), is Riemann integrable, then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi\left(X_{n:i}^{(1)}, \ldots, X_{n:i}^{(d)}\right) - \sqrt{n}\,\bar{\gamma} \xrightarrow{dist} N(0, \sigma^2) \tag{10.7}$$

as $n \to \infty$ where $\bar{\gamma}$ is defined as in Theorem 10.6 and

$$\sigma^2 = 2\sum_{j=1}^{d}\int_0^1\int_0^y x(1-y)\psi_j(x)\psi_j(y)dxdy$$
$$+2\sum_{1\le j<k\le d}\int_0^1\int_0^1 [G_{j,k}(x,y)-xy]\psi_j(x)\psi_k(y)dxdy$$

where $G_{j,k}(x,y) = F_{j,k}\left(F_j^{-1}(x), F_k^{-1}(y)\right)$.

If we arrange the random variables in a tabular form as below:

$$\begin{array}{ccccc} X_1^{(1)} & \cdots & X_1^{(j)} & \cdots & X_1^{(d)} \\ X_2^{(1)} & \cdots & X_2^{(j)} & \cdots & X_2^{(d)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{array}$$

In Theorem 10.6, we just require *i.i.d.* for each column and no further assumptions are made on how the columns are related. We need a stronger assumption in Theorem 10.7, namely, the rows are *i.i.d.* random vectors. Interestingly, the variance of the limiting normal only depends on the 2-dimensional marginal distributions.

We collect some remarks about the various conditions on the function ϕ.

Remark. (i) Conditions (C1) and (C2), in general, are easy to verify.
A sufficient condition for condition (C4) to hold is that the second partial derivatives of ψ are continuous, and there exist K and $\alpha > 0$ such that $|\psi_{j,k}(y_1,\ldots,y_d)| \le K\sum_{\ell=1}^{d}|\psi_{j,k}(y_\ell,\ldots,y_\ell)|$ and $|\psi_{j,k}(x,\ldots,x)| = O([x(1-x)]^{-\alpha})$ for $1 \le j,k \le d$ as $x \to 0+$ or as $x \to 1-$.

(ii) By a compactness argument, Condition (C1) is shown to be equivalent to

(C1') For any $c \in (0, \frac{1}{2})$, $\lim_{\delta\to 0}\omega(c,\delta) = 0$ where

$$\omega(c,\delta) := \sup|\psi(x_1,\ldots,x_d) - \gamma(y)| \tag{10.8}$$

and the supremum is taken over all $c < x_1, x_2, \ldots x_d, y < 1-c$ such that $|x_i - y| < \delta$.

(iii) Let $\psi(x,y) = [(x+y)/2]^\alpha[1-(x+y)/2]^\alpha$. It is easy to verify that for $\alpha > -1/4$, ψ satisfies conditions (C3).

References

1. Babu GJ, Bai ZD, Choi KP, Mangalam V (2009) Asymptotics of the average of functions of order statistics. J Bernoulli Soc (In press)
2. Bhattacharyya GK, Johnson RA (1977) Statistical concepts and methods, Wiley, New York
3. Billingsley P (1995) Probability and measure, Wiley, New York

Chapter 11
On Tests of Fit Based on Grouped Data

Sherzod M. Mirakhmedov and Saidbek S. Mirakhmedov

Abstract The problem of testing the goodness-of-fit of a continuous distribution for a set of n observations grouped into N equal probability intervals is considered. It is assumed that $N \to \infty$ as $n \to \infty$. Let $\eta_1, \ldots, \eta_N$ be the numbers of observations in the intervals. We show that within the class of tests based on statistics of the form $f(\eta_1) + \cdots + f(\eta_N)$ the classical chi-square test is optimal in terms of the Pitman's and the Kalenberg's intermediate asymptotic efficiencies but it is much inferior to tests satisfying Cramer condition in terms of the Kalenberg's strong intermediate and the Bahadur's exact asymptotic efficiencies. For the chi-square statistic, a probability of large deviation result, likely to be of its own interest, is proved.

11.1 Introduction

The classical goodness-of-fit problem of statistical inference is to test whether a sample has come from a given population. Specifically, we consider the problem of testing the goodness of fit of a continuous distribution F to a set of n observations grouped into N equal probability intervals. A large class of the tests based on statistics are of the general form

$$R_N^f = \sum_{k=1}^{N} f(\eta_k), \tag{11.1}$$

where $\eta_1, \ldots, \eta_N$ are the numbers of observations in the intervals. In the sequel the statistics R_N^f is called f-statistics, a test based on f-statistics is called f-tests. Three most popular cases of f-statistics are

S.M. Mirakhmedov (✉)
Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi 23460, Pakistan
e-mail: shmirakhmedov@yahoo.com

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_11,

$$X_N^2 = \sum_{k=1}^{N} \eta_k^2$$

is called the chi-square statistic,

$$\Lambda_N = \sum_{k=1}^{N} \eta_k \ln \eta_k$$

is called the likelihood ratio statistic,

$$\mu_r = \sum_{m=1}^{N} \Uparrow \{\eta_m = r\}$$

is a number of intervals containing exactly r observations, here $\Uparrow \{A\}$ is the indicator of the event A. Particularly μ_0 is known empty boxes statistic (see, for instance, [16]).

We consider the problem of testing the null hypothesis H_0: $f(x) = F'(x) = 1$, $0 < x < 1$, versus the sequence of alternatives

$$H_1 : f(x) = 1 + d\,\delta(n)\,l(x), \tag{11.2}$$

where constant $d > 0$, $l(x)$ is a known continuous function on [0,1] such that

$$\int_0^1 l(x)dx = 0, \quad \int_0^1 l^2(x)dx = 1, \quad \text{and} \quad \int_0^1 |l(x)|^3\,dx < \infty,$$

$\delta(n) \to \delta < \infty$ will be chosen so that the power for a f-test of size ω has a limit in $(\omega,1)$. The problem of testing H_0 against H_1 is called problem (H_0, H_1).

Holst [4], Ivchenko and Medvedev [6], and Gvanceladze and Chibisov [3] have shown that for the problem (H_0, H_1), for $\delta(n) = n^{-1/2}$ the power of the f-tests tends to the significance level as $n \to \infty$ whenever $N \to \infty$; hence such alternatives can not be detected by f-tests. In [20], see also [30], it was pointed out that if we let $N \to \infty$ then a sequence of alternatives that convergence to uniform must be in the form (11.2) with $\delta(n) = (n\alpha)^{-1/4}$, where $\alpha = n/N$, in order to keep the power bounded away from the significance level and unity, hence the f-tests do not discriminate alternatives (11.2) with $\delta(n) = o\left((n\alpha)^{-1/4}\right)$. This is poor in comparison with other tests based on empirical distribution functions, for example the Kolmogorov–Smirnov and Cramer–von Mises tests, which can detect similar alternatives at distance $O\left(n^{-1/2}\right)$. On the other hand not always we need to consider the alternatives converging to hypothesis with the extreme rate of $O\left(n^{-1/2}\right)$. Moreover, concerning the choice of the number of groups in chi-square test there is a well-known result by [17] stating that the optimal number N increase with n as $N = O\left(n^{2/5}\right)$. Hence it is unnatural to keep fixed number of groups when number

of observations goes to infinity. Nevertheless, because of $\delta(n) = (n\alpha)^{-1/4}$ the number of groups should go to infinity slowly in order to discriminate alternatives converging to the hypothesis with the rate close to $O\left(n^{-1/2}\right)$.

We are concerned with asymptotic results when $N = N(n) \to \infty$ as $n \to \infty$. This case has been studied intensively by many authors. We refer to [1, 4, 6, 9–11, 15, 18–20, 27, 29, 30]. For R_N^f, [7, 8, 21, 36] prove: the central limit theorem under mild condition together with Berry–Esseen bound, Edgeworth type asymptotic expansion with exact formula for the first three terms, specified for statistics X_N^2, Λ_N and μ_r, and Cramer type large deviation result under Cramer condition $E \exp\{H\,|f(\xi)|\} < \infty$, $\exists H > 0$, where ξ is the Poisson(α) r.v. with $\alpha = n/N$. We refer also to [34] where under Cramer condition the Chernoff type large deviation result for R_N^f was proved. Note that the statistics Λ_N and μ_r satisfy Cramer condition whereas the chi-square statistic X_N^2 does not. Nevertheless [30] obtain asymptotic results for Chernoff type large deviation probabilities of X_N^2 and Λ_N. These probabilistic results have been used to study asymptotic efficiencies (AE) of the f-tests. Details of the corresponding results are as follows.

There are two basic ways of comparison of tests. One of them is in principle based on asymptotic analysis of the power of the tests. A test having maximal power within a class of tests under consideration is called asymptotic most powerful (AMP) test. AMP test is also first order efficient test, and it may be not unique. In such a case an asymptotic behavior, as $n \to \infty$, of the difference in powers of two AMP tests is of interest; this situation gives rise to the concepts of second order efficient tests. The definition of the second order efficient test adapted to our problem is given later.

Another method of comparison of two tests of the same level is based on comparison of the number of observations needed to get same asymptotic power, when number of observations increases. If we have two tests with corresponding numbers of observations n_1 and n_2, then the limit of ratio n_2/n_1 is called the asymptotic relative efficiency (ARE) of test 1 with respect to (w.r.t.) test 2. To investigate AE of a test we consider that ratio where n_1 and n_2 corresponds to that test and the AMP test respectively. AE of a test depends on three parameters: the level ω_n, the power β_n and the alternative H_1, which may depend on n. When letting n go to infinity, three concepts arise: Pitman approach when $\omega_n \to \omega > 0$, $H_1 \to H_0$ in such rate that $\beta_n \to \beta \in (\omega, 1)$; Bahadur approach when $\omega_n \to 0$, $\beta_n \to \beta \in (0, 1)$, H_1 is fixed, i.e., does not approach the hypothesis; Kallenberg intermediate approach when $\omega_n \to 0$, $\beta_n \to \beta \in (0, 1)$, $H_1 \to H_0$ but more slowly than in Pitman case. Optimality of a test can be expressed by first order efficiency, which means that $n_2(\omega, \beta, H_1)/n_1(\omega, \beta, H_1)$ converges to 1, where the limit is taken according to the efficiency concept involved. The problem of finding of the limit of ratio $n_2(\omega, \beta, H_1)/n_1(\omega, \beta, H_1)$ is a very difficult problem as it usually reduces to finding of the ratio of what is called slopes of the tests under consideration, see, for instance, [5, 28].

Let ξ be the Poisson(α) r.v. with $\alpha = n/N$; P_i, E_i, Var_i stands for the probability, expectation and variance under H_i; $A_{i,N}$ and $\sigma_{i,N}^2$ stands for the asymptotic

value of $\mathrm{E}_i R_N^f$ and $Var_i R_N^f$, $i = 0, 1$, respectively. Put

$$g(\xi) = f(\xi) - Ef(\xi) - \gamma(\xi - \alpha), \gamma = \alpha^{-1} cov(f(\xi), \xi) \tag{11.3}$$

$$\sigma^2(f) = Var\, g(\xi) = Var\, f(\xi) \left(1 - corr^2\left(f(\xi), \xi\right)\right).$$

From Theorem 2 of [21] it follows that if

$$\kappa_N = \frac{E\,|g(\xi)|^3}{\sqrt{N}\sigma^3(f)} \to 0, \tag{11.4}$$

as n and $N \to \infty$, then

$$\nabla_N = \sup_x \left| P_i \left\{ R_N^f < x\sigma_{iN} + A_{iN} \right\} - \Phi(x) \right| = O\left(\kappa_N + \frac{1}{\sqrt{n}}\right), \tag{11.5}$$

where $\Phi(x)$ is the standard normal distribution function.

Remark 11.1. The statistics X_N^2 and Λ_N satisfy the condition (11.4) if and only if $n\alpha \to \infty$. But for the statistic μ_r (11.4) is valid under additional conditions for r and α, namely, if $\alpha \to 0$ and $n\alpha \to \infty$ then (11.4) satisfies for μ_r with $0 \le r \le 2$; if α is far away from zero then (11.4) is still true for μ_r, $r \ge 0$, if $\alpha - \ln N - r \ln\ln N \to \infty$.

In what follows we will assume that $n\alpha \to \infty$.

Asymptotically Most Powerful Test. It is known, see [4,7,8], that under alternative (11.2) with $\delta(n) \to 0$

$$A_{0N} = NEf(\xi), \sigma_{0N}^2 = N\sigma^2(f), \sigma_{1N}^2 = \sigma_{0N}^2(1 + o(1)), \tag{11.6}$$

$$x_N(f) \stackrel{def}{=} \left(A_{1N}(f) - A_{0N}(f)\right)/\sqrt{N}\sigma(f) = \sqrt{\frac{n\alpha}{2}}\delta^2(n)\rho(f,\alpha)d^2\,(1 + o(1))\,, \tag{11.7}$$

with $\rho(f,\alpha) = corr\left(f(\xi) - \gamma\xi, \xi^2 - (2\alpha + 1)\xi\right)$.

Let $\beta(f)$ be an asymptotic power of the f-test of a size ω. If (11.4) is satisfied then (11.5) and (11.7) imply

$$\beta(f) = \Phi\left(\sqrt{\frac{n\alpha}{2}}\delta^2(n)\rho(f,\alpha)d^2 - u_\omega\right), u_\omega = \Phi^{-1}(1 - \omega). \tag{11.8}$$

Hence the functional $\rho(f,\alpha)$ plays a key rule in determining of the asymptotic quality of the f-test. Its meaning is clarified by the following (see Lemma 1 by [8])

$$\rho(f,\alpha) = \lim corr_0\left(R_N^f, X_N^2\right). \tag{11.9}$$

Hence $|\rho(f,\alpha)| \le 1$, and $|\rho(f,\alpha)| = 1$ for any α only for chi-square test. The equality (11.8) means that f-test does not detect alternatives (11.2) with $\delta(n) = o\left((n\alpha)^{-1/4}\right)$.

In (11.2), let $\delta(n) = (n\alpha)^{-1/4}$. Then for the problem (H_0, H_1) the chi-square test is AMP within class of f-tests for any α. Nevertheless if $\alpha \to 0$ or $\alpha \to \infty$ then there exist other AMP tests also, because in these cases we may have $\rho(f,\alpha) \to 1$. For example, if $\alpha \to 0$ and $\Delta^2 f(0) \neq 0$, where operator $\Delta f(x) = f(x+1) - f(x)$, then

$$\rho(f,\alpha) = 1 - \frac{\alpha}{6}\frac{\Delta^3 f(0)}{\Delta^2 f(0)} + O\left(\alpha^2\right);$$

if $\alpha \to \infty$ for the statistic Λ_N

$$\rho(f,\alpha) = 1 - \frac{1}{6\alpha}\left(1 + o(1)\right).$$

Second Order Asymptotic Efficiency. Comparison of the AMP tests, when $\alpha \to 0$ or $\alpha \to \infty$, can be based on the notion of the second order efficiency. Set $\vartheta = d^2/\sqrt{2} - u_\omega$, where ω is a size of f-test and u_ω from (11.8), $\beta_n(R_N^f;\vartheta)$ stands for the power of the f-test of a size ω. Ivchenko and Mirakhmedov [7] show the following asymptotic expansion of the power $\beta_n(X_N^2;\vartheta)$ of chi-square test

$$\beta_n(X_N^2;\vartheta) = \Phi(\vartheta) + \varepsilon_n(X_N^2;\vartheta)\left(1 + o(1)\right),$$

where

$$\varepsilon_n(X_N^2;\vartheta) = \frac{\exp\left\{-\vartheta^2/2\right\}}{\sqrt{2\pi n\alpha}}\left(\frac{1-\vartheta^2}{3\sqrt{2}} + \frac{\vartheta d^2}{2} + \sqrt{2}S_1\left(\vartheta\sqrt{\frac{n\alpha}{2}} + \frac{n}{2}\right)\right),$$

if $\alpha \to 0$,

$$\varepsilon_n(X_N^2;\vartheta) = \frac{\exp\left\{-\vartheta^2/2\right\}}{\sqrt{2\pi N}}\left(1 - \frac{\vartheta d^2}{\sqrt{2}}\right),$$

if $\alpha \to \infty$.

Here $S_1(x) = -\{x\} + 0.5$, $\{x\}$ denotes the fractional part of x. The function $S_1(x)$ is well known in the theory of asymptotic expansion of the cumulative distribution function of lattice random variables (r.v.'s). It arises here because the chi-square statistic X_N^2 is a lattice random variable with span equal to 2.

Definition. The AMP f-test is called second order asymptotic efficient (SOAE) with respect to chi-square test, if its power has the asymptotic expansion

$$\beta_n(R_N^f,\vartheta) = \Phi(\vartheta) + \varepsilon_n(f;\vartheta)\left(1 + o(1)\right),$$

with $\varepsilon_n(f;\vartheta) \to 0$ and $\varepsilon_n(f;\vartheta) = \varepsilon_n(X_N^2;\vartheta)\left(1 + o(1)\right)$.

The conclusion of [7] is as follows. Let $\alpha \to 0$, then there exist SOAE f-tests only if $\alpha = O\left((n\alpha)^{-1/2}\right)$, i.e., $n = O\left(N^{3/4}\right)$. For example empty cells test based on μ_0 is SOAE, but likelihood ratio test is not SOAE. If $\alpha \to \infty$, then SOAE does not exist.

Pitman Efficiency. Under certain regularity conditions (see, for example, [2]), the efficacy of a test based on statistic, say V, is given by $e(V) = \mu_V^2/\sigma_V^2$. Here μ_V and σ_V^2 are the mean and variance of the limiting normal distribution under the sequence of alternatives when the test statistics has been normalized to have limiting standard normal distribution under the hypothesis. In such a situation, the Pitman's ARE of one test with respect to another is the ratio of their efficacies.

Because of (11.6) and (11.8) the alternatives (11.2) with $\delta(n) = (n\alpha)^{-1/4}$ are form P_{alt}-the family of Pitman alternatives: f-test of a size $\omega_n(f) \to \omega > 0$ and $\rho(f, \alpha)$ is far away from zero has the power (see (1.8))

$$\beta_n(f) \to \beta(f) = \Phi\left(2^{-1/2} d^2 \rho(f, \alpha) - u_\omega\right) \in (\omega, 1).$$

For the efficacy $e(f)$ of the f-test we have

$$e(f) = x_N^2(f) = 2^{-1} d^4 \rho^2(f, \alpha).$$

Hence the Pitman asymptotic efficiency of f-test is determined by functional $\rho(f, \alpha)$ and within class of f-tests the chi-square test is asymptotic most efficient (AME) in Pitman sense; the Pitman efficiency of f-tests goes down as number of intervals N increases for a given sample size n. These results have been proved by [4, 6, 20, 30].

Bahadur Efficiency. Another extreme family of alternatives is B_{alt}-Bahadur (as well as Hodges–Lehman) family of alternatives when alternatives do not approach the hypothesis, i.e., $\delta(n)$ is constant. The Bahadur's AE of f-tests in the family B_{alt} have been developed by Ronzhin (1984) who showed, for a certain subclass of f-tests, that whenever $\beta_n(f) \to \beta(f) < 1$, $-n^{-1} \log \omega_n(f)$ converges to limit (which is called slope of the f-test to the alternative) of specifies the Bahadur AE of f-test. This limit is determined by the logarithmic rate deviations probabilities (for deviation of order $O(\sqrt{N})$) under H_0, which require restrictive Cramer's condition (see (11) below) on the test statistics. In particular, this condition excludes the chi-square statistic. A comparative analysis of chi-square test's Bahadur efficiency relative to the likelihood ratio test was carried out by [30]. They showed that likelihood ratio test is much more Bahadur efficient than the chi-square test, in contrast to their relative Pitman efficiency. In a similar setup, the ARE, as $\alpha \to \infty$, of the chi-square and likelihood ratio tests were studied by [15] based on the analysis of probabilities of large deviations of various orders.

Kallenberg Intermediate Efficiency. The situation when $\delta(n) \to 0$ but slower than that in the P_{alt} give rise to the Kallenberg's intermediate family of alternatives of three types: K_{all}-family of alternatives (11.2) with $\delta(n) \to 0$, $\delta(n)(n\alpha)^{1/4} \to \infty$; $K_{1/6}$-subfamily of K_{all} with $\delta(n) = o\left((n\alpha^2)^{-1/6}\right)$ if $\alpha \geq 1$, and $\delta(n) =$

$o\left((n\alpha)^{-1/6}\right)$ if $\alpha < 1$; $K_{\sqrt{\log}}$-subfamily of K_{all} with $\delta(n)=O\left(n\alpha)^{-1/4}\log^{1/4} N\right)$. Actually such division of family of intermediate alternatives becomes from probability of large deviations results presented below in Sect. 11.3 and because of relation (11.7). Following the logic of Bahadur's approach, intermediate AE (IAE) (between Pitman and Bahadur settings) of f-tests for the family of intermediate alternatives (K_{all}, $K_{1/6}$ and $K_{\sqrt{\log}}$) can be measured by the logarithmic rate of decrease of the test size when the power is fixed. Therefore, [8] consider as a measure of the performance of f-test, the asymptotic value of a slope

$$e_n^{\omega}(R_N^f) = -\log P_0\left\{R_N^f \geq NA_{1N}(f)\right\}. \tag{11.10}$$

ARE of one test relative to another is defined as the ratio of its asymptotic slopes. For Pitman's alternatives this is equal to the Pitman's asymptotic relative efficiency, whereas for intermediate alternatives it is related to the IARE in weak sense introduced by [5]. Ivchenko and Mirakhmedov [8] extended this above said result on the Pitman's and the Bahadur's efficiencies properties of the chi-square test: chi-square test is still optimal within the class of f-tests in the Kallenberg's weak IAE (i.e., for the family $K_{\sqrt{\log}}$) but in the Kallenberg's strong intermediate efficiency (i.e., in the family K_{all} except $K_{1/6}$) it is much inferior to those statistics satisfying the Cramer condition, particularly to likelihood ratio test.

It follows from what has been said above, that the chi-square test is AMP, SOAE, and AME in the Pitman's and the Kallenberg's weak IAE senses, but it losses optimality property in terms of the Bahadur's and the Kallenberg's strong efficiency. For the f-tests satisfying Cramer condition AE for all range of alternatives (11.2), i.e., for family of alternatives P_{alt}, K_{all} and B_{alt} in the situation when α is far away from zero and infinity have been studied also. AE of the chi-square test in the family of alternatives $K_{1/6}$ has been an open problem.

Our main purpose here is 2-fold. On the one hand we wish to cover that existing gap in the study of AE of the chi-square test; with this aim we prove a probability of large deviation result, likely to be of its own interest. On the other hand we will extend the result of [8] regarding to AE of chi-square test in the family K_{all} except $K_{1/6}$, for the cases $\alpha \to 0$ and $\alpha \to \infty$.

11.2 Main Result

We assume $n\alpha \to \infty$. Also we continue to use the notations as in (11.3), (11.6), (11.7) and (11.10). The main conclusion of the present paper is the following.

Theorem 11.1. *1. If* $0 < c_0 \leq \alpha \leq c_1 < \infty$ *and*

$$\mathrm{E}\exp\{H\,|f(\xi)|\} < \infty, \tag{11.11}$$

for some $H > 0$*, then in the family of alternatives* K_{all}

$$\frac{e_n^{\omega}(R_N^f)}{n\alpha\delta^4(n)} = \frac{d^4}{4}\rho^2(f,\alpha)(1+o(1)).$$

2. *In the family of alternatives* $K_{1/6}$

$$\frac{e_n^{\omega}(X_N^2)}{n\alpha\delta^4(n)} = \frac{d^4}{4}(1+o(1)).$$

3. *In the family of alternatives* K_{all} *if*

$$\delta^3(n)(n\alpha)^{1/2}\log^{-1} N \to \infty, \tag{11.12}$$

then

$$\frac{e_n^{\omega}(X_N^2)}{n\alpha\delta^4(n)} = o(1).$$

This Theorem 11.1 together with the results of [7, 8, 30] implies that chi-square test is optimal within class of f-tests in Pitman AE and Kallenberg IAE senses but it is much inferior to those statistics satisfying the Cramer condition (particularly to likelihood ratio test and μ_r,$r \geq 0$, tests) in Kallenberg strong IAE under condition (11.12) and Bahadur AE senses. These covers almost all range of alternatives. The only remaining gap in the study of the intermediate efficiency of the chi-square test is in the family K_{all} with

$$c_2(n\alpha^2)^{-1/6} \leq \delta(n) \leq c_3(n\alpha)^{-1/6}\log^{1/3} N$$

for $\alpha \geq 1$,
and

$$c_2(n\alpha)^{-1/6} \leq \delta(n) \leq c_3(n\alpha)^{-1/6}\log^{1/3} N$$

for $\alpha < 1$.

Remark 11.2. An alternative approach to testing of uniformity [0,1] is to construct tests based on "spacings". Let $X_{1n} \leq X_{2n} \leq \ldots \leq X_{n,n}$ be the order statistics of the sample $X_1, X_2, \ldots, X_n$, $W_{m,n}^{(s)} = X_{ms,n} - X_{(m-1)s,n}$, $m = 1, 2, \ldots, N'$, $W_{N'+1,n}^{(s)} = 1 - X_{N's,n}$, with notation $X_{o,n} = 0$ and $X_{n+1,n} = 1$, be their s-spacings; $N' = [(n+1)/s]$, $N = N'$ if $(n+1)/s$ is an integer and $N = N'+1$ otherwise; $W = (W_{1,n}^{(s)}, \ldots, W_{N,n}^{(s)})$. The step of the spacings s may increase together with n, but $s = o(n)$.

The order statistics $X_{0n}, X_{1n}, \ldots, X_{n,n}, X_{n+1,n}$ divide interval [0,1] to $s+1$ subintervals (groups), that is we again, actually, deal with method of grouping data. In contrast to above considered method here the ends of intervals are random and we are using, for a statistical procedure, the length of intervals instead of frequencies of intervals. Most common among tests based on spacings are tests based on statistics of the form

$$R_N^f(W) = \sum_{m=1}^{N} f\left((n+1)W_{m,n}^{(s)}\right).$$

AE properties here alike to those of f-tests (the step of spacings s plays the role of α). For example: such tests can detect alternatives (11.2) with $\delta(n) = (ns)^{-\varepsilon}$, $\varepsilon \in (0, 1/4]$, AMP test for the alternatives (11.2) with $\delta(n) = (ns)^{-1/4}$ is the Greenwood's test based on statistic

$$G_N^2 = \sum_{k=1}^{N} \left((n+1)W_{k,n}^{(s)}\right)^2.$$

While considerable attention has been devoted in literature to type $R_N^f(W)$ statistics, we are not in position to give here all the details of existing results. Reader can find detailed information, applications, and references, for example, in papers by [11, 13, 23–25, 32, 33, 37]. We wish only to refer to [9–11] where the Pitman's ARE of chi-square test and Greenwood test with $s = [\alpha]$ were studied. They have shown that if $s \to \infty$ then these two tests have same Pitman efficiency, but for fixed s (that corresponds to the case $1 \le \alpha < c < \infty$) spacings tests are preferable to comparable chi-square procedures. From Theorem 2.1 and results of [26] it follows that the same results are still true for Kallenberg IAE of chi-square test and spacings based tests.

Proof of Theorem 11.1 For proof of *Part* 1, we refer to [8]. To *prove of Part* 2 we note that

$$\begin{aligned} e_n^{\omega}(R_N^f) &= -\log P_0\left\{R_N^f \ge x_N(f)\sigma(f)\sqrt{N} + Nf(\xi)\right\} \\ &= -\log P_0\left\{\frac{R_N^f - Nf(\xi)}{\sigma(f)\sqrt{N}} \ge \sqrt{\frac{n\alpha}{2}}\delta^2(n)\rho(f,\alpha)d^2\,(1+o(1))\right\} \end{aligned}$$

because of (11.10). Therefore, Part 2 follows from Theorem 11.1 of Sect. 11.3, using the facts that $\rho(f,\alpha) = 1$ for the chi-square statistic and that $-\log \Phi(-x) = 2^{-1}x^2(1+o(1))$, as $x \to \infty$. □

Proof of Part 3. We use an approach similar to that of [30]. We will write $\eta \sim B(k, p)$ to mean that a r.v. η has the Binomial distribution with parameters n and p, $0 < p < 1$; and $\nu(n) = \left[\alpha + \sqrt{\alpha + d^2 n\alpha\delta^2(n)}\right] + 1$ where $[a]$ is an integer part of a. We have

$$\begin{aligned} P\left\{X_N^2 > NA_{1N}\right\} &= P\left\{X_N^2 - n(\alpha+1) > x_N(f)\sqrt{2n\alpha}\right\} \\ &= P\left\{\sum_{m=1}^{N}\left((\eta_m - \alpha)^2 - \alpha\right) > d^2 n\alpha\delta^2(n)\,(1+o(1))\right\} \\ &= P\left\{\sum_{m=2}^{N}\left((\eta_m - \alpha)^2 - \alpha\right) \ge 0 \Big/ \eta_1 = \nu(n)\right\} P\left\{\eta_1 = \nu(n)\right\} \end{aligned}$$

$$= P\left\{\sum_{m=1}^{N-1}\left((\hat{\eta}_m-\alpha)^2-\alpha\right)\geq 0\right\} P\left\{\eta_1=v(n)\right\}. \qquad (11.13)$$

Here $\hat{\eta}_m \sim Bi\left(n-v(n),(N-1)^{-1}\right)$. Put $\tilde{\alpha}=(n-v(n))/(N-1)$. It easy to see that $v(n))/n=\left(N^{-1}+d\delta(n)N^{-1/2}\right)(1+o(1))$ and $\tilde{\alpha}=\alpha\left(1+O\left(N^{-1}+d\delta(n)\right.\right.$ $\left.\left.N^{-1/2}\right)\right)$. We have

$$\begin{aligned}
P\left\{\sum_{m=1}^{N-1}\left((\hat{\eta}_m-\alpha)^2-\alpha\right)\geq 0\right\} &\geq P\left\{\sum_{m=1}^{N-1}(\hat{\eta}_m-\hat{\alpha})^2\geq (N-1)\alpha\right\}\\
&= P\left\{\sum_{m=1}^{N-1}\left((\hat{\eta}_m-\hat{\alpha})^2-\hat{\alpha}\right)\geq (v(n)-\alpha)\right\}\\
&= P\left\{\frac{\sum_{m=1}^{N-1}\left((\hat{\eta}_m-\hat{\alpha})^2-\hat{\alpha}\right)}{\sqrt{2(n-v(n))^2/(N-1)}}\geq d\delta(n)+o(1)\right\}\\
&\geq c>0, \qquad (11.14)
\end{aligned}$$

because of $(n-v(n))\tilde{\alpha}=n\alpha(1+o(1))\to\infty$, and hence the Central Limit Theorem for chi-square statistic $\sum_{m=1}^{N-1}(\hat{\eta}_m-\hat{\alpha})^2$ is valid, see above Remark 11.1.

Set $g(x,p)=x\log(x/p)+(1-x)\log((1-x)/(1-p))$, $x\in(0,1)$ and $p\in(0,1)$. The following Lemma 11.1 is quoted from [30]. □

Lemma 11.1. *If $\xi\sim Bi(k,p)$ then for integer kx*

$$P\left\{\xi=kx\right\}\geq 0.8\left(2\pi kx(1-x)\right)^{-1/2}\exp\left\{-kg(x,p)\right\}.$$

Note that $\eta_1\sim B(n,N^{-1})$. Therefore due to Lemma 11.1 we have

$$\begin{aligned}
P\left\{\eta_1=v(n)\right\} &\geq c\left[v(n)\left(1-v(n)n^{-1}\right)\right]^{-1/2}\exp\left\{-v(n)\log(\alpha^{-1}v(n))\right.\\
&\qquad \left.-n(1-n^{-1}v(n))\log\frac{1-v(n)n^{-1}}{1-N^{-1}}\right\}\\
&\geq c\left(v(n)\right)^{-1/2}\exp\left\{-v(n)\log(\alpha^{-1}v(n))\right\}.
\end{aligned}$$

Hence

$$\begin{aligned}
-\frac{\log P\left\{\eta_1=v(n)\right\}}{n\alpha\delta^4(n)} &\leq c\frac{\log v(n)+v(n)\log(\alpha^{-1}v(n))}{n\alpha\delta^4(n)}\\
&\leq c\frac{\alpha+\delta(n)\sqrt{n\alpha}}{n\alpha\delta^4(n)}\log\frac{v(n)}{\alpha}\\
&\leq c\left[\frac{1}{n\delta^4(n)}+\frac{1}{\delta^3(n)\sqrt{n\alpha}}\right]\log N=o(1), \qquad (11.15)
\end{aligned}$$

because of (11.12)*. The Part 3 follows from* (11.13),(11.14) *and* (11.15)*.*

11.3 Large Deviation Results

We recall from the notations of Sect. 11.2: ξ is Poisson (α) r.v., $\alpha = n/N$, $\gamma = \alpha^{-1} cov\,(f(\xi), \xi)$,

$$g(\xi) = f(\xi) - Ef(\xi) - \gamma(\xi - \alpha)\,, \sigma^2(f) = Var\, g(\xi)$$
$$= Var f(\xi)\left(1 - corr^2\,(f(\xi), \xi)\right). \tag{11.16}$$

The following result on the large deviation probability of the chi-square statistic stated in Theorem 11.2, is likely to be of its own interest.

Theorem 11.2. *If $\alpha \geq 1$ then for all $x \geq 0$ and $x = o\left((n/\alpha)^{1/6}\right)$, if $\alpha < 1$ then for all $x \geq 0$ and $x = o\left((n\alpha)^{1/6}\right)$ one has*

$$P\left\{X_N^2 > x\sqrt{2n\alpha} + n(\alpha + 1)\right\} = (1 - \Phi(x))\,(1 + o(1))\,.$$

Proof. Let $\xi_1, \xi_2, \ldots$ be independent r.v.'s with common Poisson (α) distribution. Also let $\mathcal{C}_k(\xi)$ be a cumulant of the kth order of a r.v. ξ and

$$S_N = \sum_{m=1}^{N} (\xi_m - \alpha)\,, X_{1,N}^2 = \sum_{m=1}^{N} (\eta_m - \alpha)^2\,, \tilde{X}_{1,N}^2 = \sum_{m=1}^{N} (\xi_m - \alpha)^2\,. \square$$

Lemma 11.2. *For any fixed $k \geq 2$ and large enough n*

$$\mathcal{C}_k(X_{1,N}^2) = \mathcal{C}_k(\tilde{X}_{1,N}^2)\,(1 + o\,(1)).$$

Proof. It is well known that $\mathcal{L}((\mu_1, \ldots, \eta_N)) = \mathcal{L}((\xi_1, \ldots, \xi_N)/S_N = 0)$, where $\mathcal{L}(X)$ denotes the distribution of the random vector X. Hence

$$E(X_{1,N}^2)^k = E\left((\tilde{X}_{1,N}^2)^k / S_N = 0\right). \tag{11.17}$$

On the other hand $E\left((X_{1,N}^2)^k e^{i\tau S_N}\right) = E\left\{e^{i\tau S_N} E\left((X_{1,N}^2)^k / S_N\right)\right\}$. Integrating w.r.t. τ both side of this equality over interval $[-\pi, \pi]$, and taking into account (11.17) we have

$$E\left(X_{1,N}^2\right)^k = d_n \int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} E\left(\tilde{X}_{1,N}^2\right)^k \exp\left\{i\tau \frac{S_N}{\sqrt{n}}\right\} d\tau,$$

where

$$d_n = \left(2\pi\sqrt{n}\, P\{S_N = 0\}\right)^{-1} = \frac{1}{2\pi\sqrt{n}} \frac{n! e^n}{n^n}.$$

Hence, putting $\tilde{\xi}_m = \xi_m - \alpha$ we have

$$E\left(X_{1,N}^2\right)^k = d_n \sum_{l=1}^{k} {\sum_{k}}' {\sum_{l}}'' \int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} E\left[\tilde{\xi}_{j_1}^{2k_1} \cdot \ldots \tilde{\xi}_{j_l}^{2k_l} \exp\left\{i\tau \frac{S_N}{\sqrt{n}}\right\}\right] d\tau, \tag{11.18}$$

where $\sum_k'$ is the summation over all l-tuples $(k_1, \ldots, k_l)$ with positive integer components such that $k_1 + \cdots + k_l = k$; $\sum_l''$ is the summation over all l-tuples $(j_1, \ldots, j_l)$ with components not equal of each others and such that $j_m = 1, 2, \ldots, N; m = 1, 2, \ldots, l$.

Putting $S_{N,l} = \sum_{i=1}^{l} \tilde{\xi}_{j_i}$ we have

$$\begin{aligned}
&\int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} E\left[\tilde{\xi}_{j_1}^{2k_1} \cdot \ldots \cdot \tilde{\xi}_{j_l}^{2k_l} \exp\left\{i\tau \frac{S_N}{\sqrt{n}}\right\}\right] d\tau \\
&= \int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} E\left[\tilde{\xi}_{j_1}^{2k_1} \cdot \ldots \cdot \tilde{\xi}_{j_l}^{2k_l} \exp\left\{i\tau \frac{S_{N,l}}{\sqrt{n}}\right\}\right]\left[E \exp\left\{i\tau \frac{S_N - S_{N,l}}{\sqrt{n}}\right\}\right. \\
&\quad \left. - \exp\left\{-\frac{\tau^2}{2}\left(1 - \frac{l}{N}\right)^{-1}\right\}\right] d\tau + \int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} \exp\left\{-\frac{\tau^2}{2}\left(1 - \frac{l}{N}\right)^{-1}\right\} \\
&\quad \times E\left[\tilde{\xi}_{j_1}^{2k_1} \cdot \ldots \cdot \tilde{\xi}_{j_l}^{2k_l} \left(\exp\left\{i\tau \frac{S_{N,l}}{\sqrt{n}}\right\} - 1\right)\right] d\tau + E\left[\tilde{\xi}_{j_1}^{2k_1} \cdot \ldots \cdot \tilde{\xi}_{j_l}^{2k_l}\right] \\
&\quad \times \int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} \exp\left\{-\frac{\tau^2}{2}\left(1 - \frac{l}{N}\right)^{-1}\right\} d\tau \stackrel{def}{=} J_1 + J_2 + J_3.
\end{aligned} \tag{11.19}$$

We have

$$\left| E \exp\left\{\frac{i\tau\tilde{\xi}_m}{\sqrt{n}}\right\} \right| = \exp\left\{-2\alpha \sin^2 \frac{\tau}{2\sqrt{n}}\right\} \le \exp\left\{-\frac{2\alpha}{n\pi^2}\tau^2\right\}, \tag{11.20}$$

since $\sin^2 u/2 \ge u^2/\pi^2$, $|u| \le \pi$. Put $\Delta_n = 3N^{-1/2} + (n\alpha)^{-1/2}$, then $\sqrt{n} Var^{3/2} \xi_m / E\,|\xi_m|^3 \ge \Delta_n^{-1}$. Taking into account this inequality and using Assertion 11.4 and (11.20) by some algebra we have

$$\begin{aligned}
|J_1| \le E\left[\tilde{\xi}_{j_1}^{2k_1} \cdot \ldots \cdot \tilde{\xi}_{j_l}^{2k_l}\right] &\left\{ \int_{|\tau| \le (6\Delta_n)^{-1}} \left| E \exp\left\{i\tau \frac{S_N - S_{N,l}}{\sqrt{n}}\right\}\right.\right. \\
&\left.\left. - \exp\left\{-\frac{\tau^2}{2}\left(1 - \frac{l}{N}\right)^{-1}\right\}\right| d\tau \right. \\
&+ \int_{\pi\sqrt{n} \ge |\tau| \ge (6\Delta_n)^{-1}} \left(\left| E \exp\left\{i\tau \frac{S_N - S_{N,l}}{\sqrt{n}}\right\}\right| \right.
\end{aligned}$$

$$+\exp\left\{-\frac{\tau^2}{2}\left(1-\frac{l}{N}\right)^{-1}\right\}\Bigg)d\tau\Bigg\} \le C_4\Delta_n E\left[\tilde{\xi}_{j_1}^{2k_1}\cdot\ldots\cdot\tilde{\xi}_{j_l}^{2k_l}\right]. \tag{11.21}$$

Lemma 11.3. *For any integer $v \ge 2$ one has*

$$E\tilde{\xi}_m^v = v!\sum_{l=1}^{[v/2]} c_{l,v}\alpha^l, \tag{11.22}$$

with

$$0 < c_{l,v} < 1,\ l = 1,2,\ldots,[v/2]. \tag{11.23}$$

Proof. We have $Ee^{i\tau\tilde{\xi}} = \exp\left\{\alpha\left(e^{i\tau}-1\right)\right\}$. Applying here the Bruno's formula we find

$$E\tilde{\xi}^v = v!\sum\alpha^{k_2+\ldots+k_v}\prod_{m=2}^{v}\frac{1}{k_m!(m!)^{k_m}} = v!\sum_{l=1}^{[v/2]}c_{l,v}\alpha^l$$

where $\sum$ is summation over all non-negative $k_2,\ldots,k_v$ such that $2k_2+\ldots+vk_v = v$ and $l = k_2+\ldots+k_v$, $c_{l,v} = \sum\prod_{m=2}^{v}\frac{1}{k_m!(m!)^{k_m}}$. Lemma 11.3 follows.

In particular, from Lemma 11.3 it follows that $E\tilde{\xi}^{2k+1} \le C_1 E\tilde{\xi}^{2k}$ and $E\tilde{\xi}^{2k+2} \le C_2k^2\alpha E\tilde{\xi}^{2k}$. Using this fact after some algebra we obtain

$$\begin{aligned}|J_2| &\le \frac{1}{2n}E\left[\tilde{\xi}_{j_1}^{2k_1}\cdot\ldots\cdot\tilde{\xi}_{j_l}^{2k_l}S_{N,l}^2\right]\int_{-\infty}^{\infty}\tau^2\exp\left\{-\frac{\tau^2}{2}\left(1-\frac{l}{N}\right)^{-1}\right\}d\tau\\ &\le C\frac{\alpha k^2 l + k^2l^2}{n}E\left[\tilde{Z}_{j_1,s}^{2k_1}\cdot\ldots\cdot\tilde{Z}_{j_l,s}^{2k_l}\right].\end{aligned} \tag{11.24}$$

It is obvious that

$$J_3 = \sqrt{2\pi}E\left[\tilde{\xi}_{j_1}^{2k_1}\cdot\ldots\cdot\tilde{\xi}_{j_l}^{2k_l}\right]\left(1+O\left(\frac{l}{N}\right)\right). \tag{11.25}$$

Apply (11.21), (11.24) and (11.25) in (11.19) to get

$$\int_{-\infty}^{\infty}E\left(\tilde{\xi}_{j_1}^{2k_1}\cdot\ldots\cdot\tilde{\xi}_{j_l}^{2k_l}\exp\left\{i\tau\frac{S_N}{\sqrt{n}}\right\}\right)d\tau$$

$$= \sqrt{2\pi}\left(1+O\left(\frac{k^2}{\sqrt{n}}+\frac{\alpha k^3}{n}+\frac{k}{N}+\frac{1}{\sqrt{n\alpha}}\right)\right)E\left[\tilde{\xi}_{j_1}^{2k_1}\cdot\ldots\cdot\tilde{\xi}_{j_l}^{2k_l}\right]. \tag{11.26}$$

Using Stirling's formula it is easy to see that $d_n = (2\pi)^{-1/2}\left(1+O(n^{-1})\right)$. Inserting this and (11.26) into relation (11.18) we have

$$E\left(X_{1,N}^2\right)^k = E\left(\tilde{X}_{1,N}^2\right)^k \left(1 + O\left(\frac{\alpha k^3}{n} + \frac{1}{\sqrt{N}} + \frac{1}{\sqrt{n\alpha}}\right)\right). \tag{11.27}$$

Lemma 11.2 follows from Assertion 11.3, see Appendix, and (11.27), because $\mathcal{C}_k(X_N^2) = \mathcal{C}_k(X_{1,N}^2)$.

Let $\alpha \geq 1$. Due to Lemma 11.3 and Stirling's formula we have: for $k \geq 3$

$$\left|E\left(\tilde{\xi}_m^2 - E\tilde{\xi}_m^2\right)^k\right| \leq 2^k E\tilde{\xi}_m^{2k} \leq (2k)!2^k \sum_{l=1}^{k} \alpha^l \leq (k!)^2 \frac{2^k(2k)!}{(k!)^2} k\alpha^k$$

$$\leq (k!)^2 2^{3k+1} k\alpha^k \leq (k!)^2 \left(2^{10}k^{1/(k-2)}\alpha\right)^{k-2} Var\tilde{\xi}_m^2 \leq (k!)^2 \left(2^{12}\alpha\right)^{k-2} Var\tilde{\xi}_m^2.$$

because $Var\tilde{\xi}_m^2 = 2\alpha^2 + \alpha$. Therefore by the Assertion 11.1 with $\xi = \tilde{\xi}_m^2 - E\tilde{\xi}_m^2$, we have

$$\left|\mathcal{C}_k\left(\tilde{\xi}_m^2\right)\right| = \left|\mathcal{C}_k\left(\tilde{\xi}_m^2 - E\tilde{\xi}_m^2\right)\right| \leq (k!)^2 \left(2^{12}\alpha\right)^{k-2} Var\tilde{\xi}_m^2, k \geq 3.$$

Hence

$$\left|\mathcal{C}_k\left(\tilde{X}_N^2\right)\right| \leq (k!)^2 \left(2^{12}\alpha\right)^{k-2} N Var\tilde{\xi}_1^2, k \geq 3,$$

because as $\tilde{X}_{1,N}^2$ is a sum of i.i.d. r.v.'s. Apply this and Lemma 11.1 to get: for any fixed $k = 3, 4, \ldots$ and enough large n

$$\left|\mathcal{C}_k\left(X_N^2\right)\right| \leq (k!)^2 \left(2^{10}\alpha\right)^{k-2} \frac{Var\tilde{\xi}_1^2}{2\alpha(\alpha+1)} 2N\alpha(\alpha+1) \leq (k!)^2 \left(2^{12}\alpha\right)^{k-2} VarX_{1,N}^2,$$

since $Var\xi^2/2\alpha(\alpha+1) < 1$. Thus r.v. $X_{1,N}^2$ satisfy the Statulevicius condition (S_ν) with $\nu = 1$ and $\Delta = 2^{13}\alpha$, see Appendix. Theorem 11.1 follows from the Assertion 11.3 with $\xi = X_N^2$, $\sigma = \sqrt{2N\alpha(\alpha+1)}$ and $\Delta = 2^{13}\alpha$.

Let $\alpha < 1$. Then

$$\left|E\left(\tilde{\xi}_m^2 - E\tilde{\xi}_m^2\right)^k\right| \leq (2k)!2^k \sum_{l=1}^{k} \alpha^l \leq (k!)^2 \frac{2^k(2k)!}{(k!)^2} k\alpha \leq (k!)^2 2^{12(k-2)} Var\tilde{\xi}_m^2.$$

Hence in this case $\left|\mathcal{C}_k\left(\tilde{X}_N^2\right)\right| \leq (k!)^2 2^{12(k-2)} N Var\tilde{\xi}_1^2$ and $\left|\mathcal{C}_k\left(X_N^2\right)\right| \leq (k!)^2 2^{12(k-2)} VarX_{1,N}^2$. So r.v. $X_{1,N}^2$ satisfy the Statulevicius condition (S_ν) with $\nu = 1$and $\Delta = 2^{13}$. Theorem 11.2 follows from the Assertion 11.3 with $\xi = X_N^2$, $\sigma = \sqrt{2N\alpha(\alpha+1)}$ and $\Delta = 2^{13}$.

Appendix

Let ξ and $\xi_1, \xi_2, \ldots$, be i.i.d. r.v.'s with $E\xi = 0$, $Var\xi = \sigma^2 > 0$, $\mathcal{C}_k(\xi)$ and $\alpha_k(\xi)$ be respectively, the cumulant and moment of kth order of the r.v. ξ. The following two conditions play an essential role in the theory of large deviations (see [35]).

Bernstein's condition (B_ν): there exist constants $\nu \geq 0$ and $B > 0$ such that $|\alpha_k(\xi)| \leq (k!)^{\nu+1} B^{k-2}\sigma^2$, for all k =3,4,....

Statulevicius condition(S_ν): there exist constants $\nu \geq 0$ and $\Delta > 0$ such that $|\mathcal{C}_k(\xi)| \leq (k!)^{\nu+1} \Delta^{(k-2)}\sigma^2$, for all k =3,4,....

Assertion 11.1 ([35]). If ξ satisfies condition (B_ν) then it also satisfies the condition (S_ν) with $\Delta = 2B$.

Assertion 11.2. Suppose that r.v. ξ depending on a parameter Δ satisfies condition (S_1)(i.e., $\nu = 1$). Then $P\{\xi > x\sigma\} = \Phi(-x)\,(1 + o(1))$ for all $x \geq 0$ and $x = o\left((\Delta/\sigma)^{-1/3}\right)$.

Assertion 11.2 is Lemma 2.3 with $\nu = 1$ by [35].

Remark. The condition (S_ν) presented by [35] has the form $|\mathcal{C}_k(\xi/\sigma)| \leq (k!)^{\nu+1}$ $\Delta^{-(k-2)}$, then in the Assertions 11.1 and 11.2 one should replace $\Delta = \sigma/2B$ by $\Delta = 2B$ and $x = o\left(\Delta^{1/3}\right)$ by $x = o\left((\Delta/\sigma)^{-1/3}\right)$, respectively. The slightly different formulation presented here, of the condition (S_ν) is more convenient. For example, Assertion 11.1 now is easily understood because the following well-known relation (see, for instance [35, p. 15]).

Assertion 11.3.

$$\mathcal{C}_k(\xi) = k! \sum (-1)^{m_1+m_2+\ldots+m_k-1} (m_1+m_2+\ldots+m_k-1)! \prod_{l=1}^{k} \frac{1}{m_l!}\left(\frac{\alpha_l(\xi)}{l!}\right)^{m_l}$$

where $\sum$ is summation over all non-negative integer $m_1, m_2, \ldots, m_k$ such that $m_1 + 2m_2 + \ldots + km_k = k$

Assertion 11.4 ([31]). For all t: $|t| \leq \sigma^3\sqrt{n}/6\beta_3$, with $\beta_3 = E\,|\xi|^3$

$$\left| E \exp\left\{\frac{it}{\sigma\sqrt{n}} \sum_{m=1}^{n} \xi_m\right\} - \exp\left\{-\frac{t^2}{2}\right\} \right| \leq C \frac{\beta_3\,|t|^3}{\sigma^3\sqrt{n}} \exp\left\{-\frac{t^2}{4}\right\}.$$

References

1. Borovkov AA (1977) On the power of chi-square test with increasing number of groups. Theory Probab Appl 22:375–379
2. Fraser DAS (1957) Nonparametric methods in statistics, Wiley, New York
3. Gvanceladze LG, Chibisov DM (1979) On tests of fit based on grouped data. In: Jurechova J (ed) Contributions to statistics: Jaroslav Hájek Memorial Volume. Academia, Prague, pp 79–89

4. Holst L (1972) Asymptotic normality and efficiency for certain goodness-of-fit tests. Biometrica 59:137–145
5. Inglot T (1999) Generalized intermediate efficiency of goodness-of-fit tests. Math Methods Stat 8:487–509
6. Ivchenko GI, Medvedev YI (1978) Decomposable statistics and verifying of tests. Small sample case. Theory Probab Appl 23:796–806
7. Ivchenko GI, Mirakhmedov SA (1991) The limit theorems for divisible statistic and efficiency of corresponding tests. Discrete Math 3:73–88
8. Ivchenko GI, Mirakhmedov SA (1995) Large deviations and intermediate efficiency of the decomposable statistics in multinomial scheme. Math Methods Stat 4:294–311
9. Jammalamadaka SR, Tiwari RC (1985) Asymptotic comparison of three tests for goodness of fit. J Stat Plan Inference 12:295–304
10. Jammalamadaka SR, Tiwari RC (1987). Efficiencies of some disjoint spacings tests relative to a χ^2 tests. In: Puri ML, Vilaplana J, Wertz W (eds) New perspectives in theoretical and applied statistics. Wiley, New York, pp 311–318
11. Jammalamadaka SR, Zhou X, Tiwari RC (1989) Asymptotic efficiencies of spacings tests for goodness of fit. Metrika, 36:355–377
12. Jammalamadaka SR, Gatto R (2006) Small sample asymptotics for higher-order spacings. In: Balakrishnan N, Castillo E, Sarabia JM (eds) Advances in distributions, order statistics, and inference. Birkhauser, Basel, pp 239–252
13. Jammalamadaka SR, Goria MN (2004) A test of goodness of fit based on Gini's index spacings. Stat Probab Lett 68:177–187
14. Kallenberg WCM (1983) Intermediate efficiency, theory and examples. Ann Stat 11:170–182
15. Kallenberg WCM (1985) On moderate and large deviations in multinomial distributions. Ann Stat 13:1554–1580
16. Kolchin VF, Sevastyanov BA, Chistyakov VP (1978) Random allocation. Wiley, New York
17. Mann HB, Wald A (1942) On the choice of the number of intervals in the application of the chi-square test. Ann Math Stat 13:306–317
18. Medvedev YI (1977) Decomposable statistics in the multinomial scheme. Theory Probab Appl 22:3–17
19. Mirakhmedov SA (1985) Estimations of the closeness to the normal law of the distribution of randomized decomposable statistics in a multinomial scheme. Theory Probab Appl 30:175–178
20. Mirakhmedov SA (1987) Approximation of the distribution of multi-dimensional randomized divisible statistics by normal distribution. Theory Probab Appl 32:696–707
21. Mirakhmedov SA (1990) Randomized decomposable statistics in the scheme of independent allocations of particles into cells. Discretnaya Math (In Russian) 2(2):97–111
22. Mirakhmedov SA (1996) Limit theorems on decomposable statistics in generalized allocation scheme. Discrete Math Appl 6:379–404
23. Mirakhmedov SM[1] (2005) Lower estimation of the remainder term in the CLT for a sum of the functions of k-spacings. Stat Probab Lett 73:411–424
24. Mirakhmedov SM (2006) Probability of large deviations for the sum of functions of spacings. Int J Math Sci 2006: 1–22, Article ID 58738
25. Mirakhmedov SM, Naeem M (2008) Asymptotic properties of the goodness-of-fit tests based on spacings. Pak J Stat 24(4):253–268
26. Mirakhmedov SM (2009) On the greenwood goodness-of-fit test. J Stat Plan Inference 140(11):3017–3025.
27. Morris C (1975) Central limit theorems for multinomial sums. Ann Stat 3:165–188
28. Nikitin YY (1995) Asymptotic efficiency of nonparametric tests. Cambridge University Press, Cambridge

[1] Former Mirakhmedov S.A.

29. Quine MP, Robinson J (1984) Normal approximations to sums of scores based on occupancy numbers. Ann Probab 13:794–804
30. Quine MP, Robinson J (1985) Efficiencies of chi-square and likelihood ratio goodness-of-fit tests. Ann Stat 13:727–742
31. Petrov VV (1975) Sum of independent random variables. Ergebnisse der Mathematik und There Grenegebiete, vol 82, 1st edn. Springer, New York
32. Pyke R (1965) Spacings. J R Stat Soc Series B, 27:395–449
33. Rao JS Sethuraman J (1975) Weak convergence of empirical distribution functions of random variables subject to perturbations and scale factors. Ann Stat 3:299–313
34. Ronzhin AF (1984) Theorem on the probability of large deviations for decomposable statistics and its statistical applications. Mathe Notes 36:603–615
35. Saulis L, Statulevicius V (1991) Limit theorems for large deviations. Kluwer, Dordrecht, p 232
36. Sirajdinov SKh, Mirakhmedov SA, Ismatullaev SA (1989) Large deviation probabilities for decomposable statistics in a multinomial scheme. Theory Probab Appl 34:706–719
37. Zhou X, Jammalamadaka SR (1989) Bahadur efficiencies of spacings tests for goodness of fit. Ann Inst Stat Math 28:783–786

Chapter 12
Innovation Processes in Logically Constrained Time Series

Christoph Möller, Svetlozar T. Rachev, Young S. Kim, and Frank J. Fabozzi

Abstract Capturing the relevant aspects of phenomena in an econometric model is a fine art. When it comes to the innovation process a trade of between a suitable process and its mathematical implications has to be found.

In many phenomena the likelihood of extreme events plays a crucial role. At the same time, classical extreme value theory is based on assumptions that cannot logically be drawn for the phenomenon in question. In this paper, we exemplify the fitness of tempered stable laws to capture both the probability of extreme events, and the relevant boundary conditions in a back-coupled system, the German balancing energy demand.

12.1 Introduction

Classical time series analysis assumes a normal distribution. This assumption is justified by the central limit theorem (CLT) which states that the infinite sum of random numbers drawn from any distribution with finite first and second moments converges to the normal distribution. The global financial crisis that began in late 2007, however, has demonstrated how problematic the assumption of normal innovations or a Gaussian copular is when it comes to evaluating risk. There is a rich literature proposing alternative modeling approaches (see, e.g., [21] and the references within).

In this paper we will analyze a non-financial seasonal autoregressive integrated moving average (SARIMA) time series whose innovation process is clearly in the domain of the CLT due to physical boundary conditions. Nonetheless, we will demonstrate normal innovations to be inadequate in this setting due to the presence of heavy-tailed distributions in the data. Furthermore, we will address the issue of

C. Möller (✉)
Department of Statistics, Econometrics and Mathematical Finance, School of Economics and Business Engineering, University of Karlsruhe and KIT, Geb. 20.12, 76131 Karlsruhe, Germany
e-mail: moeller@statistik.uni-karlsruhe.de

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_12,

robustness of the Kolmogorow–Smirnow test and the Anderson–Darling test when testing heavy-tailed distributions.

The remainder of this paper is organized as follows. Section 12.2 reviews classical linear time series analysis: the non-seasonal autoregressive integrated moving average (ARIMA) and SARIMA models. Different distribution classes for the corresponding innovation process are discussed in Sect. 12.3 along with distribution tests. Section 12.4 introduces the balancing energy demand time series and the relevant boundary conditions. In Sect. 12.5 we describe the SARIMA model used to filter the innovations that are subsequently analyzed for their distribution in Sect. 12.6. Section 12.7 concludes and summarizes our results.

12.2 Time Series Models

In econometrics, a linear time series regression can be regarded as a fundamental building block. We restrict our overview here to the case of discrete time series models. One widely known example is the autoregressive (AR) model capturing serial dependence in an analyzed time series. In this model, the realized time series x_t of a random variable X_t is generated from a noise process ϵ_t, in its standard form the Gaussian white noise process, and a regression on the past realizations of the process. One example would be Brownian motion w_t in discrete time; however, this process is not stationary and may grow without limits. Nevertheless, the process is transformed into a stationary process $(w_t - w_{t-1})$ by differencing, which in turn is expanded into the original Brownian motion by summation or integration. So Brownian motion is a simple example of an integrated stationary AR-process.

If one replaces each past realization in the AR-process by the lagged process, it is transformed into a possibly infinite regression on past realizations of the noise process, the so-called moving average (MA) process. Obviously this transformation is reversible and so any such MA-process can be transformed into a possibly infinite AR-process. Additionally, Wold's decomposition theorem states that any stationary time series may be represented by a possibly infinite MA-process, or equivalently AR-process. This explains why the combination of the autoregressive moving average (ARMA)-process has become the standard approach in time series modeling. The practical advantage over the strict AR- and MA-processes is a parsimonious usage of parameters. This approach is expanded to an even wider range of non-stationary time series the ARIMA model, when including the appropriate degree of differencing to obtain a stationary time series.

Again the consideration of parameter parsimony leads to a further specification, the SARIMA model. This seasonal ARIMA model is a serial connection of ARIMA models. Here, one ARIMA model captures a serial dependence structure over cyclical periods (e.g., 1 year), and a further ARIMA model captures the dependence on a sub-cycle horizon [1]. Using the lag-operator L and the difference-operator ∇, the model can be expressed as in (12.1):

$$\phi_p(L)\Phi_P(L^s)\nabla^d\nabla^D x_t = \theta_q(L)\Theta_Q(L^s)\epsilon_t \tag{12.1}$$

Here s is the seasonality, p and P are the orders of the AR-polynomials ϕ, Φ and q and Q are the orders of MA-polynomials θ and Θ, and d and D are the appropriate numbers of differencing to obtain stationarity.

Therefore the model is also abbreviated as $SARIMA(p,d,q) \times (P,D,Q)_s$. The crucial difference to the regular $ARIMA(p,d,q)$ model can be best seen by expanding the multiplicative model into its additive form as in (12.2):

$$\begin{aligned}&(1-\phi_1 L-\phi_2 L^2-\ldots-(\phi_s+\Phi_1)L^s-(\phi_{s+1}-\phi_s\Phi_1)L^{s+1}-\ldots)\nabla^d\nabla^D x_t\\&=(1+\theta_1 L+\theta_2 L^2+\ldots+(\theta_s+\Theta_1)L^s+(\theta_{s+1}+\theta_s\Theta_1)L^{s+1}+\ldots)\epsilon_t.\end{aligned} \tag{12.2}$$

So as an example, a $SARIMA(1,0,0) \times (1,0,0)_{12}$ is equivalent to an $ARIMA$ $(13,0,0)$ with the constraint $\phi_{13} = -\phi_1\phi_{12}$.

We conclude this section with a brief overview of model identification techniques that were established by [1]. The first condition a time series needs to fulfill in order to adapt a linear model is stationarity. This condition is fulfilled if all roots of the polynomial $\phi(L)$ lie outside the unit circle. Consequently, the hypothesis of a root of value one is tested (i.e., the unit root test). Should this hypothesis be rejected, the time series may be regarded as stationary, otherwise a higher order of differencing has to be considered. The work of [6] lead to the augmented Dickey–Fuller test (ADF-test) as an unbiased unit root test.

Once the appropriate degree d of differencing is selected, inspection of the sample autocorrelation function (SACF) and partial autocorrelation function (SPACF) will help to select the order of the AR and MA polynomials. The SACF will cut off at the lag q of a $MA(q)$-process, while the SPACF will cut off at the lag p of an $AR(p)$-process. Unfortunately, the SACF will only gradually tail off for an AR-process as will the SPACF for a MA-process. Consequently, an $ARMA$-process will lead to a tailing off of both SACF and SPACF. Additionally, because it is calculated on a limited sample, this method is also prone to statistical fluctuations. Nonetheless, sudden drops in SACF and SPACF give an indication of appropriate choices for the parameters p and q of an $ARMA$-process.

Finally model selection criteria such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC) should be used as a guide to avoid the pitfalls of overparameterized modeling.

12.3 Innovations

The main purpose of modeling time series is to identify and understand underlying characteristics and then be able to apply this understanding to predict the future. While a specification of the appropriate parameters of a SARIMA$(p,d,q) \times (P,D,Q)_s)$ captures information on the location parameter of a future realization, the identification of the distribution of the innovation process ϵ_t is crucial in understanding the risk that goes along with anticipating this realization.

Among all distributions, the class of α-stable distributions stands out as the distributional class any appropriately scaled sum of independent and identically distributed (i.i.d.) random variables will converge to. In the context of time series modeling, this property is of particular interest as the innovations in each time step are considered to be the sum of many independent events from the same set of plausible events (i.e., the sum of i.i.d. random variables).

The normal distribution is the most well-known member of the class of α-stable distributions, and for this distribution the convergence property is known as the CLT. Relaxing the assumption of existing first and second moments, the CLT is generalized. The limiting distribution is now not generally the normal distribution, but, instead, a member of the class of α-stable distributions. This is known as the general central limit theorem (GCLT). In the following we will refer to α-stable distributions as non-Gaussian to distinguish the class from its most well-known member. In general, for α-stable distributions there exists no probability density function in closed form and therefore it is expressed by its characteristic function as given by (12.3):

$$\begin{aligned}\phi_{\text{stable}}(u;\alpha,\sigma,\beta,\mu) &= E[e^{iuX}] \\ &= \begin{cases} \exp\left(i\mu u - |\sigma u|^{\alpha}\left(1 - i\beta(\text{sign } u)\tan\frac{\pi\alpha}{2}\right)\right), & \alpha \neq 1 \\ \exp\left(i\mu u - \sigma|u|\left(1 + i\beta\frac{2}{\pi}(\text{sign u})\ln|\text{u}|\right)\right), & \alpha = 1, \end{cases}\end{aligned} \tag{12.3}$$

where

$$\text{sign t} = \begin{cases} 1, & t \geq 0 \\ 0, & t = 0 \\ -1, & t \leq 0. \end{cases}$$

In this parameterization, the four parameters $(\alpha, \beta, \sigma, \mu)$ have the following domain and interpretation:

- α: the index of stability or the shape parameter, $\alpha \in (0, 2)$
- β: the skewness parameter, $\beta \in [-1, +1]$
- σ: the scale parameter, $\sigma \in (0, +\infty)$
- μ: the location parameter, $\mu \in (-\infty, +\infty)$

The α-stable distribution is a very rich class of distributions that can capture even extreme asymmetric and heavy-tailed structural conditions. However, in some problems the implication of infinite variance, or in the case of $\alpha \leq 1$ infinite mean and variance, is too strong an assumption. At the same time the empirical distribution might clearly show excess kurtosis or skewness. For these problems the class of tempered stable distributions has been proposed. (See [12,16], and [15]). In an adaption of the α-stable distribution, the tempered stable distribution has the characteristic function given by (12.4) with one extra parameter.

$$\begin{aligned}\phi_{CTS}(u;\alpha,C,\lambda_+,\lambda_-,m) = \exp(&ium - iuC\Gamma(1-\alpha)(\lambda_+^{\alpha-1} - \lambda_-^{\alpha-1}) \\ &+ C\Gamma(-\alpha)((\lambda_+ - iu)^{\alpha} - \lambda_+^{\alpha} + (\lambda_- + iu)^{\alpha} - \lambda_-^{\alpha})).\end{aligned} \tag{12.4}$$

Here m and C determine the location and scale as do μ and σ in the α-stable distribution. However, the skewness is parameterized by λ_+ and λ_-, at the same time allowing for a faster than α-stable decay in the tails. So a tempered stable distribution allows for the same flexibility at the center of the distribution as the α-stable distributions, combined with finite first and second moments. Therefore, the tempered stable distribution is in the domain of the CLT, though convergence to the Gaussian limit may be very slow due to the almost α-stable distribution in the center. So tempered stable distributions are a natural choice whenever the problem suggests a distribution that is prone to extreme events, but at the same time other considerations forbid infinite moments; that is, whenever the rate of convergence to the limiting Gaussian distribution at the infinite sum is too slow to justify a Gaussian model at the finite time horizon of interest.

The classical tempered stable (CTS) distribution has been introduced under different names in the literature, including the truncated Levy flight by [14], the KoBoL distribution by [2], and the CGMY distribution by [4].

Having introduced the tempered stable distribution, we conclude this section with a review of methods to justify the introduction of additional parameters into a model by specifying a non-Gaussian innovation process. One widely used graphical method is the inspection of the quantile-quantile (QQ)-plot. Figure 12.1 shows an example of the quantiles of the fitted Gaussian-distribution versus the empirical quantiles of an observed innovations time series. Ideally, the QQ-plot would be a straight line of unit slope, and any deviation from this line is an indication of the

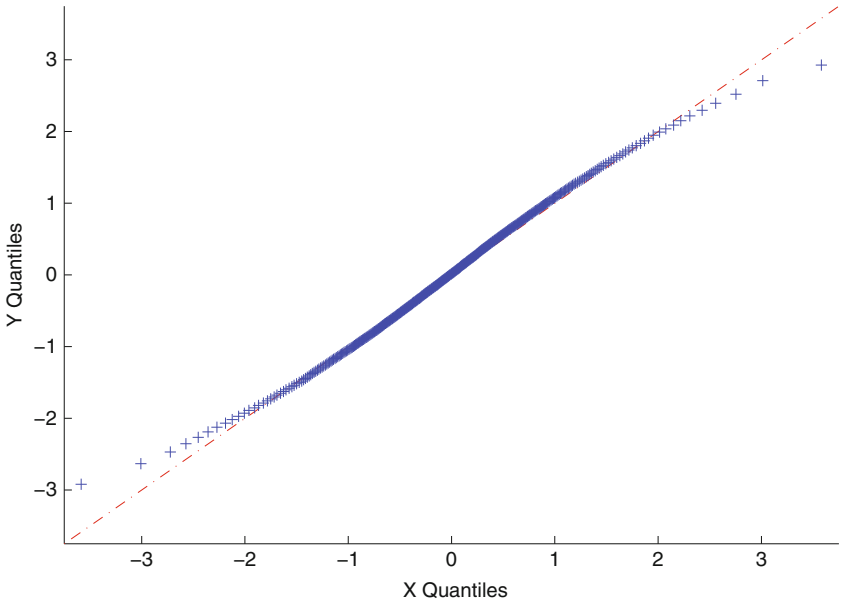


Fig. 12.1 QQ-plot of heavy-tailed data under the Gaussian hypothesis

Table 12.1 Goodness-off-fit test and their statistic

Test	Statistic
KS	$\sqrt{n}\sup\left\lvert F_{emp.}(x) - F_{th.}(x)\right\rvert$
AD	$\sqrt{n}\sup\left\lvert \frac{F_{emp.}(x)-F_{th.}(x)}{\sqrt{F_{th.}(x)(1-F_{th.}(x))}}\right\rvert$
CvM	$n\int_{-\infty}^{\infty}(F_{emp.}(x) - F_{th.}(x))^2 dF_{th.}(x)$
AD^2	$n\int_{-\infty}^{\infty}\left(\frac{F_{emp.}(x)-F_{th.}(x)}{F_{th.}(x)(1-F_{th.}(x))}\right)^2 dF_{th.}(x)$

chosen distribution not adequately describing the innovation time series. One should keep in mind that because it is based on a finite sample, the empirical quantiles will never stretch out to arbitrarily small or large quantiles.

Another less subjective method is testing the proposed distribution with a goodness-of-fit test. All the tests calculate a test statistic (see Table 12.1) to test the hypothesis of the innovations time series to be a realization of the fitted distribution. These statistics can then be compared to critical values and p-values can be calculated.

The Kolmogorov–Smirnov (KS) test is known for its critical values to be independent of the tested distribution. Like the KS-test, the Anderson–Darling (AD) test is based on the supremum of the difference of the theoretical and empirical CDF, however, a weight assigned to each point assigns more weight to the tails of the distribution. In the case of the α-stable distribution, this feature is obviously of particular importance.

Both the KS-test and AD-test have to be used with caution when testing α-stable or tempered stable distributions. We illustrate this drawback in the following example. Suppose the underlying distribution is the CTS-$(0.3, 1, 0.5, 0.7, 0)$ shown in Fig. 12.2. There is an almost vertical step in the CDF at the location parameter m. Because of the dominance of extreme events, the mean of a sample drawn from a heavy-tailed distribution will only converge slowly to the true mean of zero. Consequently, the steep increase is very likely to be dislocated in the empirical CDF. In combination with the almost vertical increase at the location parameter, this may result in large deviances between the empirical and theoretical CDF and a rejection of the distribution in the KS-test. This argument might even hold true for the AD-test where the weights are minimal at the location parameter. This behavior is particularly undesirable when analyzing innovation distributions where the true mean of zero is known and the location parameter is of minor importance. We therefore suggest a flexibility in adjusting the sample mean before a KS-test or an AD-test.

In contrast to the KS-test and AD-test, the Cramer–von Mises (CvM) test and the AD^2-test are based on the area between the theoretical and empirical CDF (see Table 12.1). The main advantage of these statistics is that they incorporate information about the total sample, and, in particular, are insensitive to a slight dislocation of the empirical CDF. The AD^2-test introduces a weighting scheme that focuses the test on the tails of the distribution. One major disadvantage of the CvM-test, AD^2-test and AD-test is that critical values are dependent on the analyzed distribution. One feasible method of obtaining critical values in these cases is to use Monte Carlo-based

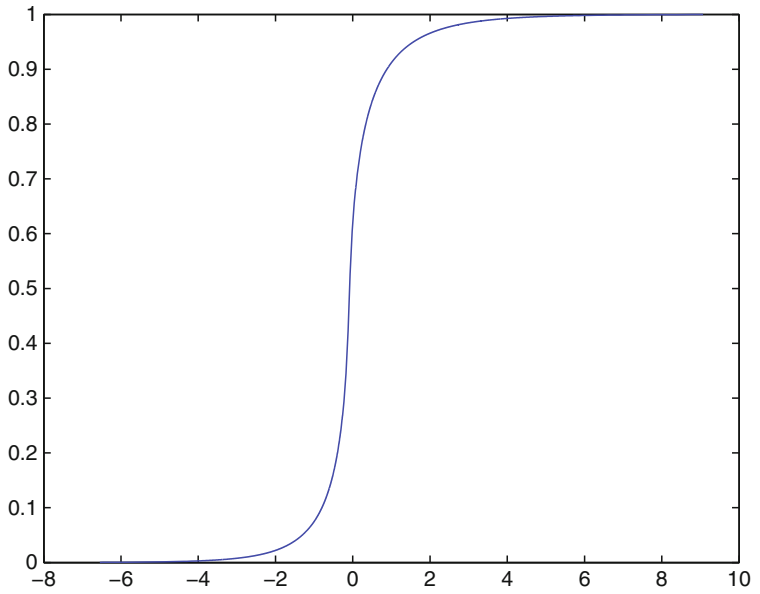


Fig. 12.2 CDF of CTS-(0.3, 1, 0.5, 0.7, 0)

simulation based on the estimated parameter set (see [5]). Therefore, the tests are not available in common software, and in turn results are less transparent.

12.4 Market Setting and Data

Electricity is a special commodity as it is practically not storable. Consequently, in electricity grids supply and demand sides have to be in exact equilibrium at all times, otherwise a blackout will result. This task of active balancing is performed by the transmission system operator (TSO) who activates bids of up- or down-regulating energy to maintain equilibrium. In this context, it is important to distinguish between regulatory energy and balancing energy. Regulatory energy is contracted prior to the actual balancing action by the TSO to allocate the resources to be able to balance the grid. Balancing energy is the energy the TSO will settle with balancing responsible entities (BR) who caused a disturbance after the balancing action. The TSO performs this correction by calling up appropriate regulatory energy bids contracted prior to the actual balancing.

It is technically not feasible to balance disturbances of arbitrary sizes as the blackouts in northeast North America in 2003 or the blackout in central Europe in 2006 demonstrate. So balancing energy demand is constrained around zero; however, no exact boundary can be specified. In general, balancing energy has to compensate all unpredictable supply and demand shocks such as a power plant

outage or a football match going into overtime. Such shocks are unknown in advance and not predictable over a time horizon of a day or even more. Within the German electricity market design, balancing energy may however also be used as a substitute for other energy contracts as long as this position is neither obvious nor excessive (see [3]). This incentive leads to pronounced patterns and serial dependence (see [19] and [18]).

In this paper we concentrate on the governing innovation distribution of balancing energy demand in Germany. To uncover this distribution, we therefore have to filter from the data predictable components. The analytical model used here is described in [18]. For the German market two factors are identified. These factors are the gradient of grid load and an arbitrage incentive. Next we restate some of the results.

The gradient effect is modeled by the out of sample average, which in turn can be modeled by (12.5) using the load ($L(t)$) as a factor.

$$q(\nabla L(t)) = q \cdot (\bar{L}_q(t) - \bar{L}_h(t)), \quad q = 0.42 \tag{12.5}$$

This model describes the quarter hourly pattern with a R^2 of 0.87. The effect is fully attributable to the discrepancy between the step function of load changes imposed by the traded contracts on the day-ahead market, and a rather smoothly changing load in reality. Such effects cannot be observed in most of the other European markets, because the settlement periods of the day-ahead market and the balancing energy market are one and the same in these markets (see [9]).

As for the second factor, there exist arbitrage incentives ($I(t), I_{tec}(t)$) to substitute electricity trades on the day-ahead market with strategic positions in the balancing energy market or avoid technical difficulties in plant operation. This incentive is driven by the day-ahead prices exceedence of a common price level and can be modeled as in (12.6). The model captures dependence and saturation effects, and is compatible with a functional relation that is constant over time as the parameter estimates and R^2-values in Table 12.2 demonstrate.

$$h(I(t), I_{tec}(t)) = a \cdot \left(\frac{2}{1 + b \cdot e^{-c \cdot I(t)}} - 1 \right) + I_{tec}(t) \quad a \in \mathbb{R}\ b, c \in \mathbb{R}_+ \tag{12.6}$$

Table 12.2 Parameters and R^2 fitting to out of sample data. Source [18]

	Parameters			R^2 factor model	
Year	$a[MWh]$	$b[E]$	$c[€^{-1}]$	$I(t)$ only	$I(t)$ and $I_{tec}(t)$
2003	940.045	1.053	0.035	0.6948	0.7252
2004	901.082	1.113	0.039	0.4448	0.6170
2005	918.633	1.089	0.039	0.6069	0.7518
2006	902.485	1.072	0.038	0.8424	0.8499
2007	928.798	1.081	0.037	0.7571	0.7998
2008	884.110	1.068	0.043	0.6530	0.7025

The sample variance is reduced by 12 and 19% in the case of the gradient effect and the arbitrage incentive, respectively. This clearly demonstrates the existence of non-random predictable components in the balancing energy demand data.

The raw data are obtained from the publications of balancing energy demand the four German TSOs from January 2003 to December 2008. Namely the four sources are RWE [22], E.on [8], EnBW [7], and Vattenfall [24]. These data from four sub-zones are combined to obtain one hypothetical German zone to account for interzonal balancing effects and provide a better correspondence to the single German day-ahead market EEX.

12.5 SARIMA Model of Balancing Energy Demand

Here we focus on the analysis of the residual time series after subtracting the effects described by the analytical model given by (12.5) and (12.6). Before trying to adapt a linear time series model to the residuals, the data need to be checked for stationarity. We apply an augmented Dickey–Fuller unit root test. The null hypothesis of a unit root is rejected at a significance level below $\alpha = 0.001$ even when including the first 168 lags for the regression.

This finding is supported also by a consideration of the physical boundary conditions of the underlying data. Balancing energy demand is fulfilled by the grid operator to ensure grid balance. This energy has to be delivered physically by power stations, and so the installed capacity imposes a hard boundary. This boundary can, however, never be reached, as the response time and response capacity of power stations impose an even tighter boundary. Due to their design, powers stations cannot run on an arbitrary fraction of their designed capacity, but instead have to be operated within a certain bandwidth. Additionally, a complex system such as a power station has a considerable amount of inertia, and cannot instantaneously adapt to changes in operation. When looking at the total generation stock, although these facts do not translate into a hard boundary, the true limits will depend on the exact condition and history of all individual facilities connected to the grid. Nonetheless a limit to fluctuations the grid operator can manage always exists. So it is physically impossible for the balancing energy demand to grow to very large positive values or fall to very small negative values, but balancing energy will always be within a bandwidth around zero. Mathematically, this argument relates to a stationary time series, and the absence of unit roots. We can therefore model the data without the need of further differencing.

As a first step, an inspection of the SACF and SPACF of the residuals in Fig. 12.3 shows the presence of SARIMA effects in the data. The autocorrelation decays off with increasing lag. Additionally, this decay is disturbed at multiples of 24, indicating a seasonality of 24. This picture is supported by the partial autocorrelation function displaying a drop at lag one and 24, together with a decaying negative

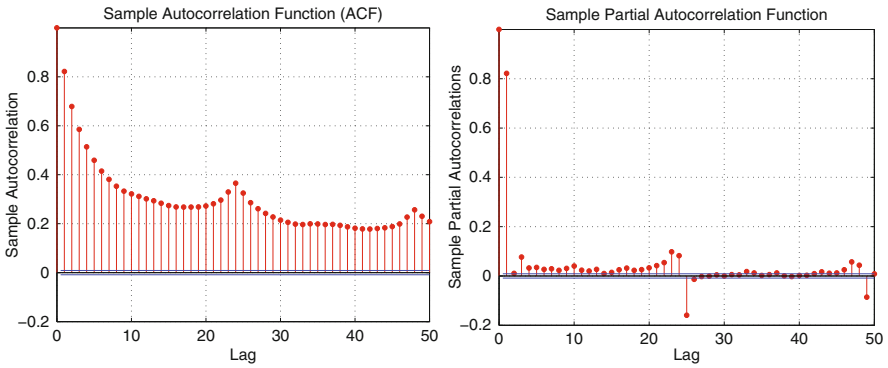


Fig. 12.3 Sample autocorrelation and partial autocorrelation

Table 12.3 Parameter estimates of SARIMA model

	SARIMA			
	$(1,0,0) \times (1,0,1)_{24}$		$(3,0,0) \times (1,0,1)_{24}$	
	Gaussian	t(ν)	Gaussian	t(ν)
ar_1	0.8185	0.8238	0.7974	0.8036
ar_3	–	–	0.0084	0.0079
ar_{24}	0.9572	0.9571	0.9419	0.9427
ma_{24}	−0.8502	−0.8529	−0.8291	−0.8328
σ	341.9658	341.9410	341.7744	342.0834
ν	–	9.1080	–	9.1396
AIC	763,170	762,080	763,060	761,970
BIC	763,200	762,110	763,090	762,000

partial autocorrelation at lags following multiples of 24. Moreover, the SPACF indicates another step at lag three. We therefore choose SARIMA$(1,0,0) \times (1,0,1)_{24}$ and SARIMA$(3,0,0) \times (1,0,1)_{24}$ as candidates for the model. Additional to the classical model with Gaussian innovations we include models with t-distributed innovations as representatives of heavy-tailed innovations in the analysis. This approach is suggested by [25] as a compromise between a heavy-tailed innovation distribution and robust parameter estimates for the SARIMA-model. Table 12.3 holds the AIC and BIC values of different specifications, including both Gaussian and t-distributed innovations. We choose the SARIMA$(1,0,0) \times (1,0,1)_{24}$ model with t-distributed innovations for two reasons. First, the AIC and BIC values indicate a preference of t-distributed innovations over the Gaussian case. Second, the ar_3 coefficient is small and, as we will discuss below, coefficients at low lags are of minor practical relevance. Note that the t-distributed innovations demonstrate the necessity of a heavy-tailed noise term in the model. This will be further investigated in Sect. 12.6.

We conclude this section with an analysis of the consistency of the model over time and a test of its forecasts. Table 12.4 reports the parameter estimates of the

Table 12.4 Parameter estimates of SARIMA model

Parameter	a_1	a_{24}	b_{24}	σ	ν
Total	0.8238	0.9571	−0.8529	341.9410	9.1080
2003	0.8522	0.9307	−0.7973	341.9410	9.4983
2004	0.7860	0.9421	−0.8184	328.8776	12.3724
2005	0.7810	0.9359	−0.8248	325.6431	10.7623
2006	0.8092	0.9455	−0.8397	336.8019	9.0538
2007	0.8067	0.9586	−0.8620	329.9785	10.1529
2008	0.7861	0.9515	−0.8541	341.9410	9.2676

model based on yearly sub-samples. The parameter estimates are very consistent with the overall model. We therefore decided to test the forecasts of the overall model rather than the individual yearly models.

The information on balancing energy demand is not continuously revealed to the market, but rather published only once a month, including the data for the preceding month. So the data for May will be available by July. We therefore do not test the one-time step forecast, as this has no practical implication in this market. Instead, we test a forecast adapted to the information revealed to the market. As a result, forecasting is performed once a month based on the information lagged 1 month (i.e., the forecast horizon is 720–1,440 lags). Because one of the TSOs also publishes the balancing demand data in its zone with a time delay of only 3 days, we also test the implications of the model on this time horizon (i.e., a forecast horizon of 72–96 lags). In both cases that we test, the sample variance is reduced by subtracting the conditional expectation. When using the monthly forecast, the variance is reduced by 3.69%; applying a 3-day forecast horizon results in a 11.22% reduction.

This additional variance reduction as compared to the analytical model in Sect. 12.4 can be decomposed into two components. For the first component there exist comparatively short-lived patterns in the data. These patterns can be understood as a linear correction term for the analytical model. The second component captures a non-zero conditional mean of the time series. Neither the gradient effect nor the arbitrage incentive described in Sect. 12.4 can explain a non-zero expectation value of the balancing energy demand over extended periods of time. Both effects will average to zero over a few cycles of their respective seasonality. However when looking at the average forecast of the SARIMA-model in Table 12.5, it is evident the SARIMA forecast does not average to zero over a few cycles. Furthermore, it is argued in [17] that this non-zero mean over extended periods of time is sufficient to influence the electricity price on the day-ahead market. Other studies of market power abuse in Germany identify the same periods as periods of abuse, which show highly negative mean SARIMA forecasts (i.e., artificial demand) in their analysis (see [10, 11, 23]).

Table 12.5 Average prediction of SARIMA model

	Average prediction					
Horizon	2003	2004	2005	2006	2007	2008
One month	−198.9537	−179.9403	−110.6992	21.8594	−118.4479	−52.1579
Three days	−372.3728	−307.2270	−214.6126	18.6222	−217.6485	−190.4806

12.6 Innovation Distribution

The parameter estimation of the SARIMA-model used in Sect. 12.5 is based on a t-distributed innovation process. Models with Gaussian innovations were disregarded based on AIC and BIC values reported in Sect. 12.5. In this section, we will further investigate the heavy-tailed innovation process.

The innovations process is of particular importance in the balancing energy market as it governs the risk involved with balancing the grid. In general, TSOs have to allocate sufficient regulatory power to be able to maintain grid operation and avoid a blackout. The capacity that is considered sufficient is usually defined by a threshold probability for a blackout (i.e., the probability of fluctuations exceeding the allocated capacity). So the more precise the quantiles of the innovations distributions are known, the more efficiently resources may be allocated.

The first step of the investigation is the QQ-plot of the innovations time series and the fitted t-distribution in Fig. 12.4. From this figure it can be seen that the t-distribution does not adequately capture the risk in the tails of the empirical distribution because the QQ-plot deviates from the diagonal.

Due to the conceptual advantage of modeling the data with a distribution in the proximity of the GCLT as discussed in Sect. 12.3, we test both the α-stable and the CTS-distribution as a more adequate model for the innovations time series. Both distributions are estimated by the Fourier inversion formula and their characteristic functions given by (12.3) and (12.4). This inversion is, in turn, numerically estimated by the fast Fourier transform (FFT) method. A more detailed description of the method is given in [20] and [13]. Table 12.6 shows the estimated parameter sets. As can be seen in Fig. 12.5, the heavy-tailed distribution captures the likelihood of extreme events more accurately than the t-distribution.

In the next step, all three distributions are compared using the goodness-of-fit tests mentioned in Sect. 12.3. The results are summarized in Table 12.7.

The p-values of the KS-test clearly indicate that the CTS-distribution describes the innovations best, as its p-value is 49 and 9 orders of magnitude greater than the p-value of t-distribution and α-stable distribution, respectively. However the p-value of the CTS-distribution is still low. As discussed in Sect. 12.3, the KS-test is responsive to small fluctuations in the location parameter, while such fluctuations are to be expected with heavy-tailed distributions. Also, the SARIMA model implies a location parameter of zero for the innovation process, so we do not need to focus on the location parameter. We therefore correct the mean of the innovations time series for such fluctuations within the 95% confidence bounds. The corresponding

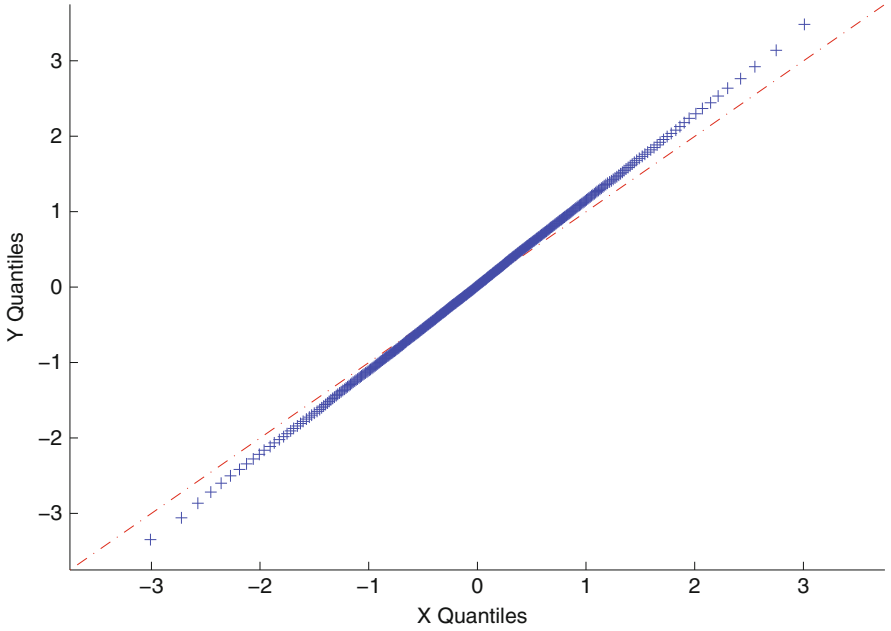


Fig. 12.4 QQ-plot t-distribution

Table 12.6 Estimated parameters of heavy-tailed distributions

Distribution	Parameters				
α-stable	α	σ	β		μ
	1.9107	0.0048	0.6711		0-fixed
CTS	α	C	λ_+	λ_-	m
	0.9122	$\frac{1}{\Gamma(2-\alpha)(\lambda_+^{\alpha-2}+\lambda_-^{\alpha-2})}$	1.4856	1.5168	0-fixed

statistics are identified by an asterisk (*). Again the CTS-distribution provides the best description of the data. Furthermore, the CTS-distribution is acceptable at a 5% significance level. The other statistics reported provide further support for selecting the CTS-distribution over both the t-distribution and the α-stable distribution.

12.7 Conclusion

The wide application of linear time series models in finance has made the shortcomings of inadequate innovation processes and correlation structures evident. Tempered stable distributions have been proposed and tested on financial time series data to overcome these problems. The tempered stable distribution is an expansion of the α-stable distribution. It combines heavy-tailed innovations over multiple timescales with finite higher moments. These properties make tempered stable

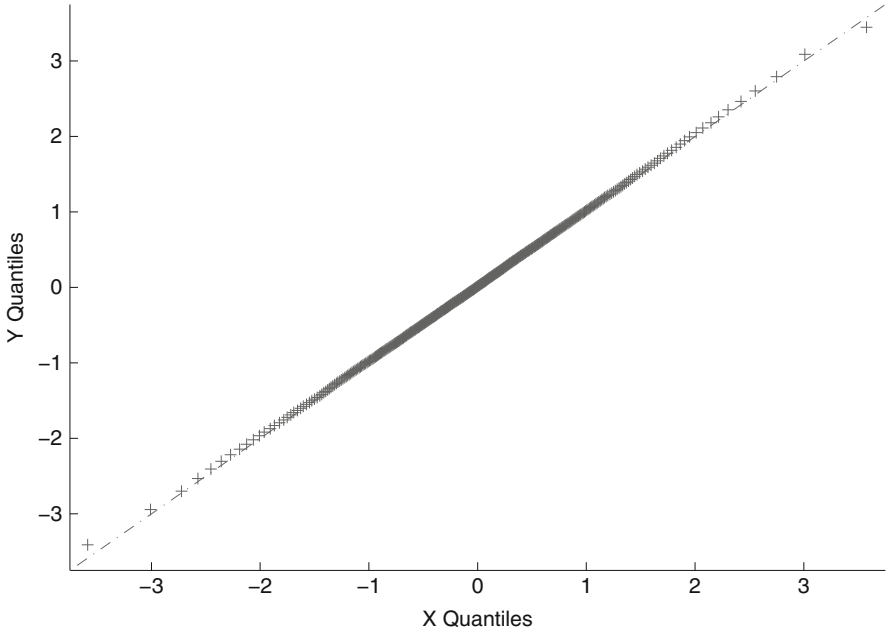


Fig. 12.5 QQ-plot CTS-distribution

Table 12.7 Goodness-of-fit statistics and p-values

	Distribution					
	t		α-stable		CTS	
Test	Statistic	p-value	Statistic	p-value	Statistic	p-value
KS	0.0331	$1.37 \cdot 10^{-50}$	0.0156	$1.75 \cdot 10^{-11}$	0.0082	0.0016
AD	0.0883	–	0.0333	–	0.0170	–
KS*	0.0316	$5.73 \cdot 10^{-46}$	0.0134	$1.40 \cdot 10^{-8}$	0.0059	0.0538
AD*	0.0854	–	0.0316	–	0.0126	–
CvM*	23.4809	–	2.2939	–	0.5144	–
AD^2*	625.3888	–	685.0225	–	611.5575	–

distributions an excellent choice for modeling phenomena dependant on extreme events, which are at the same time bounded by other considerations.

In this paper, we apply the classical tempered stable model to German balancing energy demand data and demonstrate its fitness. The CTS-distribution describes the risk of unpredictable events in the electricity grid, while at the same time capturing physical boundary conditions in the model. In the balancing energy market, these advantages can be used to allocate the resource capacity more efficiently, and thereby reduce grid tariffs. Additionally, the SARMIA model is able to separate predictable balancing energy demand from unpredictable shocks. In general, the predictable fraction of balancing energy demand can be satisfied by a wider and technically less demanding range of providers. The German market design includes

the option to trade this predictable fraction in strategic positions on the balancing market, while the reserve capacity for the unpredictable shocks is traded on the market for regulatory energy. A further development of strategic positions in the balancing energy market could therefore increase the competitiveness of the capacity reserve market. In view of the planned increase of highly fluctuating regenerative power in the German system, this advantage will become even more pronounced in the future.

Acknowledgements Svetlozar T. Rachev gratefully acknowledges research support by grants from Division of Mathematical, Life and Physical Sciences, College of Letters and Science, University of California, Santa Barbara, the DFG and the DAAD. The authors thank Prof. Gennady Samorodnitsky for his help in formulating the problem of TID distributions and for his fruitful comments and suggestions.

References

1. Box G, Jenkins GM (1970) Time series analysis: Forecasting and control. Holden-Day, San Francisco
2. Boyarchenko SI, Levendorskiĭ SZ (2000) Option pricing for truncated Lévy processes. Int J Theor Appl Finance 3:549–552
3. Bundesnetzagentur (2006) Bilanzkreisvertrag über die Führung eines Bilanzkreises. http://www.bundesnetzagentur.de
4. Carr P, Geman H, Madan D, Yor M (2002) The fine structure of asset returns: An empirical investigation. J Bus 75(2):305–332
5. Chernobai A, Rachev S, Fabozzi F (2007) Operational risk. The Frank J. Fabozzi Series. Wiley Finance, Hoboken, New Jersey
6. Dickey D, Fuller W (1979) Distribution of the estimators for autoregressive time series with a unit root. J Am Stat Assoc 74(366):427–431
7. EnBW Transportnetze (2009) Bilanzkreisabrechnung: Verfahren und preise. https://www.regelleistung.net
8. e.on Netz (2009) Bilanzkreisabweichung. https://www.regelleistung.net
9. EU (2005) Technical Annexes to the Report from the Commission on the Implementation of the Gas and Electricity Internal Market COM(2004)863 final. Commission staff working document, Commission of the European Communities
10. EU (2007) Structure and performance of six European wholesale electricity markets in 2003, 2004 and 2005. European commission dg competition report, London Economics and Global Energy Decisions
11. von Hirschhausen C, Weigt H, Zachmann G (2007) Preisbildung und Marktmacht auf den Elektrizitätsmärkten in Deutschland. Expert opinion(VIK), Universität Dresden EE2
12. Kim YS, Rachev ST, Bianchi ML, Fabozzi FJ (2008) Financial market models with Lévy processes and time-varying volatility. J Bank Finance 32:1363–1378
13. Kim YS, Rachev ST, Bianchi ML, Fabozzi FJ (2009) Computing var and avar in infinitely divisible distributions. Technical report, Chair of Econometrics, Statistics and Mathematical Finance School of Economics and Business Engineering University of Karlsruhe (http://www.statistik.uni-karlsruhe.de/download/AVaR_ID_KRBF_03-25.pdf)
14. Koponen I (1995) Analytic approach to the problem of convergence of truncated Lévy flights towards the gaussian stochastic process. Phys Rev E 52:1197–1199
15. Menn C, Rachev ST (2009) Smoothly truncated stable distributions, GARCH-models, and option pricing. Math Meth Oper Res 69:411–438

16. Menn C, Rachev ST, Fabozzi FJ (2005) Fat-tailed and skewed asset return distributions: Implications for risk management, portfolio selection, and option pricing. Wiley, Hoboken, New Jersey
17. Möller C, Rachev S, Fabozzi F (2009) Strategic deployment of balancing energy in the German electricity market. In: Proceedings of the 10th IAEE European conference energy, policies and technologies for sustainable economies. IAEE Austria
18. Möller C, Rachev S, Fabozzi F (2010) Balancing energy strategies in electricity portfolio management. Energy Econ (In press)
19. Nailis D, Ritzau M (2006) Studie zur Marktausgestalltung der Regel-Ausgleichsenergie vor dem Hintergrund des neuen EnWG. Techical report, BET
20. Nolan JP (1997) Numerical calculation of stable densities and distribution functions. Commun Stat Stochastic Models 13:759–774
21. Rachev ST, Mittnik S (2000) Stable paretian models in finance. Wiley, New York
22. RWE Transportnetz Strom (2009) Bilanzkreisabweichungen. https://www.regelleistung.net
23. Schwarz HG, Lang C (2006) The Rise in German wholesale electricity prices: Fundamental factors, exercise of market power, or both? Technical report, University of Erlangen-Nürnberg IWE
24. Vattenfall Europe Transmission (2009) Ausgleichsenergie Arbeitspreise. https://www.regelleistung.net
25. Zumbach G (2006) A gentle introduction to the RM 2006 methodology. Technical report, RiskMetrics Group

Chapter 13
Laws of Large Numbers and Nearest Neighbor Distances

Mathew D. Penrose and J.E. Yukich

Abstract We consider the sum of power weighted nearest neighbor distances in a sample of size n from a multivariate density f of possibly unbounded support. We give various criteria guaranteeing that this sum satisfies a law of large numbers for large n, correcting some inaccuracies in the literature on the way. Motivation comes partly from the problem of consistent estimation of certain entropies of f.

13.1 Introduction

Nearest-neighbor statistics on multidimensional data are of long-standing and continuing interest, because of their uses, e.g., in density estimation and goodness-of fit testing [3, 11, 21], and entropy estimation [2, 4, 9, 10]. They form a multivariate analog to the one-dimensional spacings statistics in which the work of S.R. Jammalamadaka, the dedicatee of this paper, has featured prominently. For example, [16] uses nearest neighbor balls to generalize the maximum spacings method to high dimensions and to establish consistency in estimation questions.

In the present note we revisit, extend and correct some of the laws of large numbers concerned with sums of power-weighted nearest-neighbor distances that have appeared in recent papers, notably [10, 15, 19].

Fix $d \in \mathbb{N}$ and $j \in \mathbb{N}$. Given a finite $\mathcal{X} \subset \mathbb{R}^d$, and given a point $x \in \mathcal{X}$, let $\mathrm{card}(\mathcal{X})$ denote the number of elements of $\mathcal{X}$, and let $D(x, \mathcal{X}) := D_j(x, \mathcal{X})$ denote the Euclidean distance from x to its jth nearest neighbor in the point set $\mathcal{X} \setminus \{x\}$, if $\mathrm{card}(\mathcal{X}) > j$; set $D(x, \mathcal{X}) := 0$ if $\mathrm{card}(\mathcal{X}) \leq j$. Let f be a probability density function on $\mathbb{R}^d$, and let $(X_i)_{i \in \mathbb{N}}$ be a sequence of independent random d-vectors with common density f. For $n \in \mathbb{N}$, let $\mathcal{X}_n := \{X_1, \ldots, X_n\}$. Let $\alpha \in \mathbb{R}$ and set

M.D. Penrose (✉)
Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK
e-mail: m.d.penrose@bath.ac.uk

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_13,

$$S_{n,\alpha} := \sum_{x \in \mathcal{X}_n} (n^{1/d} D(x, \mathcal{X}_n))^\alpha = \sum_{i=1}^{n} (n^{1/d} D(X_i, \mathcal{X}_n))^\alpha.$$

Certain transformations of the $S_{n,\alpha}$ have been proposed [9, 10] as estimators for certain entropies of the density f which are defined in terms of the integrals

$$I_\rho(f) := \int_{\mathbb{R}^d} f(y)^\rho dy \quad (\rho > 0).$$

For $\rho > 0$ with $\rho \neq 1$, the Tsallis ρ-entropy (or Havrda and Charvát ρ-entropy [7]) of the density f is defined by $H_\rho(f) := (1 - I_\rho(f))/(1 - \rho)$, while the Rényi entropy [17] of f is defined by $H^*_\rho(f) := \log I_\rho(f)/(1 - \rho)$.

Rényi and Tsallis entropies figure in various scientific disciplines, being used in dimension estimation and the study of nonlinear Fokker–Planck equations, fractal random walks, parameter estimation in semi-parametric modeling, and data compression (see [4] and [10] for further details and references).

A problem of interest is to estimate the Rényi and Tsallis entropies, or equivalently, the integrals $I_\rho(f)$, given only the sample $\{X_i\}_{i=1}^n$ and their pairwise distances. Let $\omega_d := \pi^{d/2}/\Gamma(1 + d/2)$ denote the volume of the unit radius Euclidean ball in d dimensions, and set $\gamma(d, j) := \omega_d^{-\alpha/d} \left(\frac{\Gamma(j+\alpha/d)}{\Gamma(j)} \right)$. This note provides sufficient conditions on the density f establishing that $\gamma(d, j)^{-1} n^{-1} S_{n,\alpha}$ converges to $I_{1-\alpha/d}(f)$ in L^1, or in L^2. In other words, since L^1 convergence implies convergence of means, we provide sufficient conditions on f guaranteeing that $\gamma(d, j)^{-1} n^{-1} S_{n,\alpha}$ is an *asymptotically unbiased and consistent estimator* of $I_{1-\alpha/d}(f)$.

13.2 Results

Two of our results can be stated without further ado.

Theorem 13.1. *Let $\alpha > 0$. Suppose the support of f is a finite union of convex bounded sets with nonempty interior, and f is bounded away from zero and infinity on its support. Then as $n \to \infty$ we have L^2 and almost sure convergence*

$$n^{-1} S_{n,\alpha} \to \omega_d^{-\alpha/d} \left(\frac{\Gamma(j + \alpha/d)}{\Gamma(j)} \right) I_{1-\alpha/d}(f). \tag{13.1}$$

Theorem 13.2. *Let $q = 1$ or $q = 2$. Let $\alpha \in (-d/q, 0)$ and suppose f is bounded. Then (13.1) holds with L^q convergence.*

For the interesting case when $\alpha > 0$ and f has unbounded support, our results require further notation. Let $|\cdot|$ denote the Euclidean norm on $\mathbb{R}^d$. For $r > 0$, define the integral

$$M_r(f) := \mathbb{E}\,[|X_1|^r] = \int_{\mathbb{R}^d} |x|^r f(x)dx,$$

and define the critical moment $r_c(f) \in [0, \infty]$, by

$$r_c(f) := \sup\{r \geq 0 : M_r(f) < \infty\}.$$

If $r < s$ and $M_s(f) < \infty$, then $M_r(f) < \infty$. Hence $M_r(f) < \infty$ for $r < r_c(f)$ and $M_r(f) = \infty$ for $r > r_c(f)$.

For $k \in \mathbb{N}$, let A_k denote the annular shell centered around the origin of $\mathbb{R}^d$ with inner radius 2^k and outer radius 2^{k+1}, and let A_0 be the ball centered at the origin with radius 2. For Borel measurable $A \subset \mathbb{R}^d$, set $F(A) := P[X_1 \in A] = \int_A f(x)dx$.

We can now state the rest of our results.

Theorem 13.3. *Let $q = 1$ or $q = 2$. Let $\alpha \in (0, d/q)$. Suppose $I_{1-\alpha/d}(f) < \infty$, and $r_c(f) > q\alpha d/(d - q\alpha)$. Then (13.1) holds with L^q convergence.*

We shall deduce from Theorem 13.3, that when $f(x)$ decays as a power of $|x|$, the condition $I_{1-\alpha/d}(f) < \infty$ is sufficient for L^1 convergence:

Corollary 13.1. *Suppose there exists $\beta > d$ such that $f(x) = \Theta(|x|^{-\beta})$ as $|x| \to \infty$, i.e., such that for some finite positive C we have*

$$C^{-1}|x|^{-\beta} < f(x) < C|x|^{-\beta}, \quad \forall x \in \mathbb{R}^d, \ |x| \geq C. \tag{13.2}$$

Suppose also that $I_{1-\alpha/d}(f) < \infty$ for some $\alpha \in (0, d)$. Then (13.1) holds with L^1 convergence.

Our final result shows that in general, the condition $I_{1-\alpha/d}(f) < \infty$ is *not* sufficient alone for L^1 convergence, or even for convergence of expectations. It can also be viewed as a partial converse to Theorem 13.3 showing, under the additional regularity condition (13.3), that when $q = 1$ the condition $r_c(f) > q\alpha d/(d - q\alpha)$ is close to being sharp.

Theorem 13.4. *Let $0 < \alpha < d$. Then (i) If $r_c(f) < \alpha d/(d - \alpha)$, and also for some $k_0 \in \mathbb{N}$ we have*

$$0 < \inf_{k \geq k_0} \frac{F(A_k)}{F(A_{k-1})} \leq \sup_{k \geq k_0} \frac{F(A_k)}{F(A_{k-1})} < \infty, \tag{13.3}$$

then $\limsup_{n\to\infty} \mathbb{E}\,[n^{-1} S_{n,\alpha}] = \infty$;

(ii) For $0 < r < \alpha d/(d - \alpha)$ there exists a bounded continuous density function f on $\mathbb{R}^d$ satisfying (13.3), such that $I_{1-\alpha/d}(f) < \infty$, but with $r_c(f) = r$ so that $\limsup_{n\to\infty} \mathbb{E}\,[n^{-1} S_{n,\alpha}] = \infty$ by part (i).

The value of the limit in (13.1) was already known (see Lemma 13.1). The contribution of the present paper is concerned with the conditions under which the

convergence (13.1) holds; in what follows we compare our conditions with the existing ones in the literature and also comment on related limit results. For conditions under which $n^{-1/2}(S_{n,\alpha} - \mathbb{E}\, S_{n,\alpha})$ is asymptotically Gaussian, we refer to [1,12,14].

Remark. (i) *Theorem 13.1.* The condition in Theorem 13.1 is a slight relaxation of condition C1 of the L^2 convergence results in [15] or [19], which assume a polyhedral support set. When the support of f is the unit cube, Theorem 2.2 of [8] gives an alternative proof of almost sure convergence in (13.1) (we remark that Theorem 2.2 of [8] contains an extraneous $\mathbb{E}$ in the left-hand side). Theorem 8.3 of [20] uses subadditivity to obtain (13.1) when the support of f is the unit cube and when $\alpha = 1$, but it does not identify the limiting constant. The convergence of means implied by Theorem 13.1 was previously obtained, under some extra differentiability conditions on f, in [5].
(ii) *Theorem 13.2.* The L^1 convergence of Theorem 13.2 improves upon Theorem 3.1 of [10], which establishes mean convergence; the L^2 convergence of Theorem 13.2 is contained in Theorem 3.2 of [10] and we include this for completeness.

(iii) *Theorem 13.3.* The condition in Theorem 13.3 corrects the condition of the corresponding result given [15], where for L^1 convergence it is stated that we need $r_c(f) > d/(d-\alpha)$; in fact we need instead the condition $r_c(f) > \alpha d/(d-\alpha)$. In the proof of Theorem 13.3 below, we shall indicate the errors in the proof in [15] giving rise to this discrepancy. This correction also applies to condition C2 in Theorem 2 of [19], the proof of which relies on the result stated in [15].

(iv) *Theorem 13.4.* The condition (13.3) holds, e.g., if $f(x)$ is a regularly varying function of $|x|$. Given (13.3) and given $I_{1-\alpha/d} < \infty$, Theorem 13.4 shows that the condition $r_c(f) \geq \alpha d/(d-\alpha)$ is necessary for L^1 convergence of $n^{-1}S_{n,\alpha}$, while Theorem 13.3 says that $r_c(f) > \alpha d/(d-\alpha)$ is sufficient. It would be of interest to try to find more refined necessary and sufficient conditions when $r_c(f) = \alpha d/(d-\alpha)$.

(v) *General ϕ.* For $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ put $S_{n,\phi} := \sum_{x \in \mathcal{X}_n} \phi(n^{1/d} D(x, \mathcal{X}_n))$. If ϕ has polynomial growth of order α, that is if there is a constant $\alpha \in (0, \infty)$ such that $\phi(x) \leq C(1 + x^\alpha)$ for all $x \in \mathbb{R}^+$, then straightforward modifications of the proofs show that under the conditions of Theorem 13.1 or Theorem 13.3 we have the corresponding L^q convergence

$$n^{-1} S_{n,\phi} \to \int_{\mathbb{R}^d} \mathbb{E}\,[\phi(D(\mathbf{0}, \mathcal{P}_{f(x)}))] f(x) dx,$$

where for all $\tau > 0$, $\mathcal{P}_\tau$ is a homogeneous Poisson point process in $\mathbb{R}^d$ having constant intensity τ, and $D(\mathbf{0}, \mathcal{P}_\tau)$ is the distance between the origin of $\mathbb{R}^d$ and its jth nearest neighbor in $\mathcal{P}_\tau$.

(vi) *Minimal spanning trees.* Given a finite $\mathcal{X} \subset \mathbb{R}^d$ and $\phi : \mathbb{R}^+ \to \mathbb{R}^+$, let

$$L_\phi(\mathcal{X}) := \sum_{e \in MST(\mathcal{X})} \phi(|e|),$$

where $\mathrm{MST}(\mathcal{X})$ denotes the edges in the graph of the minimal spanning tree on $\mathcal{X}$. Thus $L_\phi(\mathcal{X})$ is the sum of the ϕ-weighted edge lengths in the minimal spanning tree on $\mathcal{X}$. Let $q = 1$ or 2. If ϕ has polynomial growth of order α, with $\alpha \in (0, d/q)$, if $I_{1-\alpha/d}(f) < \infty$, and if $r_c(f) > q\alpha d/(d - q\alpha)$ then, as may be seen by following the proof of Theorem 13.3, the proof of Theorem 2.3(iii) of [15] in fact shows that as $n \to \infty$ we have

$$L_\phi(\mathcal{X}_n) \to \frac{1}{2} \int_{\mathbb{R}^d} \mathbb{E} \left[\sum_{e \in MST(\mathbf{0}, \mathcal{P}_{f(x)})} \phi(|e|) \right] f(x) dx,$$

where the convergence is in L^q, and where $MST(\mathbf{0}, \mathcal{P}_{f(x)})$ denotes the edges in the minimal spanning tree graph on $\mathbf{0} \cup \mathcal{P}_{f(x)}$ incident to $\mathbf{0}$, the origin of $\mathbb{R}^d$. When $q = 2$, this is new whereas for $q = 1$ and $\alpha \in (0, 1)$, this improves upon Theorem 2.3(iii) of [15], which requires $r_c(f) > \max(\alpha d/(d - \alpha), d/(d - \alpha))$.

(vii) *Non-existence of density.* If the $\{X_i\}_{i=1}^n$ fail to have a density, then normalization of $S_{n,\alpha}$ may involve exotic functions of n, including log periodic normalizations, as is the case when the $\{X_i\}_{i=1}^n$ have a Cantor distribution on $[0, 1]$; see [18].

(viii) *Comparison with [10].* The convergence of expectations corresponding to (13.1) is given as the main conclusion in Theorem 3.1 of [10]. In the case $1 - \alpha/d < 1$ of that result, it is claimed that this convergence of expectations holds without any extra conditions besides finiteness of $I_{1-\alpha/d}$. Theorem 13.4 here disproves this assertion; the argument in [10] requires that convergence in distribution implies convergence of rth moments, which is not in general true. On the other hand, Corollary 13.1 shows that if we assume $f(x)$ decays as some power of $|x|$ then finiteness of $I_{1-\alpha/d}$ is indeed a sufficient condition for convergence in L^1, and hence also convergence of expectations.

13.3 Proofs

This section provides the proofs of the results stated in the preceding section. We denote by c, C, C', and C'' various strictly positive finite constants whose values may change from line to line. The proofs of Theorems 13.1, 13.2 and 13.3 use the following result.

Lemma 13.1. *Let $q \in \{1, 2\}$ and $\alpha \in \mathbb{R}$. Suppose for some $p > q$ that $\mathbb{E}[(n^{1/d} D(X_1, \mathcal{X}_n))^{\alpha p}]$ is a bounded function of n. Then (13.1) holds with L^q convergence.*

Proof. Since D is a stabilizing functional on homogeneous Poisson point processes [15], we can apply Theorem 2.2 of [15] or Theorem 2.1 of [15] to get L^q convergence of $n^{-1} S_{n,\alpha}$ to a limit which is expressed as an integrated expectation in [15] (see (2.15) of [15]). It was shown in [19] that this limit is equal to the right hand side of (13.1) (and this is also consistent with the limiting constant in [5]). □

Proof of Theorem 13.1. Recall that we assume the support of f, namely $\text{supp}(f) := \{x \in \mathbb{R}^d : f(x) > 0\}$, is a finite union of bounded convex sets with nonempty interior, here denoted $B_1, \ldots, B_m$. Set $\lambda := \sup\{|x - y| : x \in \text{supp}(f), y \in \text{supp}(f)\}$, the diameter of the support of f. By assumption, $\lambda < \infty$. Also we assert that there is a constant $c > 0$ such that for $r \in (0, \lambda]$,

$$F(B_r(x)) \geq cr^d, \quad \forall x \in \text{supp}(f). \tag{13.1}$$

To see this, take $\delta_1 > 0$ such that for $1 \leq i \leq m$ there is a ball B_i^- of radius δ_1 contained in B_i. There is a constant $\delta_2 > 0$ such that for $1 \leq i \leq m$, if $x \in B_i$, and $r \leq \delta_1$, then the intersection of the ball of radius r centered at x with the convex hull of the union of B_i^- and x has volume at least $\delta_2 r^d$. This region is contained in B_i and (13.1) follows for $r \in (0, \delta_1]$. But then (with a different choice of c) (13.1) follows for $r \leq \lambda$. Hence, for $0 < t \leq \lambda n^{1/d}$ and with $B(x; r)$ denoting the Euclidean ball of radius r centered at x,

$$\begin{aligned} P[n^{1/d} D(X_1, \mathcal{X}_n) > t] &\leq \sup_{x \in \text{supp}(f)} P[\text{card}(\mathcal{X}_{n-1} \cap B(x; n^{-1/d} t)) < j] \\ &\leq \sum_{i=0}^{j-1} \binom{n-1}{i} (cn^{-1}t^d)^i (1 - cn^{-1}t^d)^{n-1-i} \\ &\leq C \sum_{i=0}^{j-1} t^{id} \exp(-cn^{-1}t^d (n - 1 - i)) \\ &\leq C'(1 + t^{(j-1)d}) \exp(-ct^d) \leq C'' \exp(-(c/2)t^d). \end{aligned}$$

Moreover this probability is clearly zero for $t > \lambda n^{1/d}$. Hence, for $\alpha > 0$ and $p > 2$,

$$\begin{aligned} \mathbb{E}\,[(n^{1/d} D(X_1, \mathcal{X}_n))^{\alpha p}] &= \int_0^\infty P[n^{1/d} D(X_1, \mathcal{X}_n) > u^{1/(\alpha p)}] du \\ &\leq C \int_0^\infty \exp(-(c/2)u^{d/(\alpha p)}) du \end{aligned}$$

which is finite and does not depend on n. Therefore we can apply Lemma 13.1 to get the L^2 convergence (13.1).

For almost sure convergence, we apply Theorem 2.2 of [13], where here the test function considered in that result (and denoted f there, not to be confused with the notation f as used here) is the identity function. It is well known (see [3], or Lemma 8.4 of [20]) that there is a constant $C := C(d)$ such that for any finite $\mathcal{X} \subset \mathbb{R}^d$, any point $x \in \mathcal{X}$ is the jth nearest neighbor of at most C other points of $\mathcal{X}$. Therefore adding one point to a set $\mathcal{X}$ within the bounded region $\text{supp}(f)$ changes the sum of the power-weighted jth nearest neighbor distances by at most a constant. Therefore (2.9) of [13] holds here (with $\beta = 1$ and $p' = 5$ say), and the almost sure convergence follows by Theorem 2.2 of [13]. □

Proof of Theorem 13.2. The proof depends on the following lemma. Recall that $(X_i)_{i\geq 1}$ are i.i.d. with density f. Given X_1, let V_n denote the volume of the d-dimensional ball centered at $n^{1/d}X_1$ whose radius equals the distance to the jth nearest point in $n^{1/d}(\mathcal{X}_n \setminus X_1)$, where for $r > 0$ and $\mathcal{X} \subset \mathbb{R}^d$ we write $r\mathcal{X}$ for $\{rx : x \in \mathcal{X}\}$. For all $x \in \mathbb{R}^d$, for all $n = 2, 3, \ldots$ and for all $v \in (0, \infty)$ let

$$F_{n,x}(v) := P[V_n \leq v | X_1 = x]. \tag{13.2}$$

Lemma 13.2. *If f is bounded and $\delta \in (0, 1)$, then*

$$\sup_n \mathbb{E}\, V_n^{-\delta} = \sup_n \int_{\mathbb{R}^d} \int_0^\infty v^{-\delta}\, dF_{n,x}(v) f(x) dx < \infty.$$

Proof. Since $\int_0^\infty v^{-p}\, dF(v) = p \int_0^\infty v^{-p-1} F(v) dv$ for any $p \in (0, 1)$ whenever both integrals exist (see, e.g., Lemma 1 on p. 150 of [6]), we have for all $x \in \mathbb{R}^d$

$$\begin{aligned}
\int_0^\infty v^{-\delta}\, dF_{n,x}(v) &= \delta \int_0^\infty v^{-\delta-1} F_{n,x}(v) dv \\
&\leq \int_0^1 v^{-\delta-1} F_{n,x}(v) dv + \delta \int_1^\infty v^{-\delta-1} dv \\
&= \int_0^1 v^{-\delta-1} F_{n,x}(v) dv + 1.
\end{aligned}$$

With $\tilde{B}_v(x)$ denoting the ball of volume v around x, for all $v \in (0, 1)$ we have

$$\begin{aligned}
F_{n,x}(v) &= P[V_n \leq v | X_1 = x] = 1 - P[\mathrm{card}(n^{1/d}\mathcal{X}_{n-1} \cap \tilde{B}_v(n^{1/d}x)) < j] \\
&\leq 1 - P[\mathrm{card}(n^{1/d}\mathcal{X}_{n-1} \cap \tilde{B}_v(n^{1/d}x)) = 0] \\
&= 1 - \left(1 - \int_{\tilde{B}_{v/n}(x)} f(z) dz\right)^{n-1}.
\end{aligned} \tag{13.3}$$

Since f is assumed bounded we have

$$F_{n,x}(v) \leq 1 - \exp\left((n-1)\log(1 - \|f\|_\infty v/n)\right).$$

When n is large enough, for all $v \in (0, 1)$ we have $(n-1)\log(1 - \|f\|_\infty v/n) \geq -2\|f\|_\infty v$, and so for all $x \in \mathbb{R}^d$

$$F_{n,x}(v) \leq 1 - \exp(-2\|f\|_\infty v) \leq 2\|f\|_\infty v.$$

Hence for all n large enough and all x we have $\int_0^1 v^{-\delta-1} F_{n,x}(v) dv \leq 2\|f\|_\infty \int_0^1 v^{-\delta-1} v dv$, demonstrating Lemma 13.2.

Now to prove Theorem 13.2, we choose $p > q$ such that $-1 < \alpha p/d < 0$ and invoke Lemma 13.2 to conclude $\sup_n \mathbb{E}\,[V_n^{\alpha p/d}] < \infty$. We now apply Lemma 13.1 to complete the proof of L^q convergence. □

The proof of Theorem 13.3 uses the following lemma. Recall from Sect. 13.2 the definition of the regions $A_k, k \geq 0$.

Lemma 13.3. *Let $0 < s < d$. If $r_c(f) > sd/(d-s)$, then $\sum_{k=1}^{\infty} 2^{ks} (F(A_k))^{(d-s)/d} < \infty$.*

Proof. We modify some of the arguments on p. 85 of [20]. For all $\varepsilon > 0$, by Hölder's inequality we have

$$\begin{aligned}\sum_k 2^{ks} (F[A_k])^{(d-s)/d} &= \sum_k 2^{-\varepsilon ks} (F[A_k])^{(d-s)/d} 2^{(1+\varepsilon)ks} \\ &\leq \left(\sum_k (2^{-\varepsilon ks})^{d/s} \right)^{s/d} \left(\sum_k F[A_k] (2^{(1+\varepsilon)ks})^{d/(d-s)} \right)^{(d-s)/d} \\ &\leq C(\varepsilon, s) \left(\sum_k \int_{A_k} |x|^{(1+\varepsilon)sd/(d-s)} f(x) dx \right)^{(d-s)/d}\end{aligned}$$

which, for ε small enough, is finite by hypothesis. □

Proof of Theorem 13.3. We follow the proof in [15], but correct it in some places and give more details in others. We aim to use Lemma 13.1. Since we assume $0 < \alpha < d/q$, we can take $p > q$ with $\alpha p < d$. Clearly

$$\mathbb{E}\,[(n^{1/d} D(X_1, \mathcal{X}_n))^{\alpha p}] = n^{\alpha p/d - 1} \mathbb{E}\left[\sum_{i=1}^{n} D(X_i, \mathcal{X}_n)^{\alpha p} \right] = n^{\alpha p/d - 1} \mathbb{E}\,[L^{\alpha p}(\mathcal{X}_n)], \tag{13.4}$$

where for any finite point set $\mathcal{X} \subset \mathbb{R}^d$, and any $b > 0$, we write $L^b(\mathcal{X})$ for $\sum_{x \in \mathcal{X}} D(x, \mathcal{X})^b$ (and set $L^b(\emptyset) := 0$). Note that for some finite constant $C = C(d, j)$ the functional $\mathcal{X} \mapsto L^b(\mathcal{X})$ satisfies the simple subadditivity relation

$$L^b(\mathcal{X} \cup \mathcal{Y}) \leq L^b(\mathcal{X}) + L^b(\mathcal{Y}) + C t^b \tag{13.5}$$

for all $t > 0$ and all finite $\mathcal{X}$ and $\mathcal{Y}$ contained in $[0, t]^d$ (cf. (2.2) of [20]).

As in (7.21) of [20] or (2.21) of [15] we have that

$$L^{\alpha p}(\mathcal{X}_n) \leq \left(\sum_{k=0}^{\infty} L^{\alpha p}(\mathcal{X}_n \cap A_k) \right) + C(p) \max_{1 \leq i \leq n} |X_i|^{\alpha p}. \tag{13.6}$$

In the last sentence of the proof of Theorem 2.4 of [15] it is asserted that the last term in (13.6) is not needed, based on a further assertion that one can take $C = 0$

in (13.5) here, but these assertions are incorrect. For example, if $\mathrm{card}(\mathcal{Y}) \le j$ then $L^b(\mathcal{Y}) = 0$ but $L^b(\mathcal{X} \cup \mathcal{Y})$ could be strictly greater than $L^b(\mathcal{X})$. Similarly, if $\mathrm{card}(\mathcal{X}_n \cap A_k) \le j$ then the term in (13.6) from that k is zero but the corresponding contribution to the left side of (13.6) is non-zero.

Combining (13.6) with (13.4) yields

$$\mathbb{E}\left[(n^{1/d} D(X_1, \mathcal{X}_n))^{\alpha p}\right] \le n^{(\alpha p - d)/d} \mathbb{E}\left[\sum_k L^{\alpha p}(\mathcal{X}_n \cap A_k)\right] + C(p)\mathbb{E}\left[n^{(\alpha p - d)/d} \max_i |X_i|^{\alpha p}\right]. \tag{13.7}$$

By Jensen's inequality and the growth bounds $L^{\alpha p}(\mathcal{X}) \le C(\mathrm{diam}\mathcal{X})^{\alpha p} (\mathrm{card}(\mathcal{X}))^{(d-\alpha p)/d}$ (see Lemma 3.3 of [20]), we can bound the first term in the right hand side of (13.7) by

$$C \sum_k 2^{k\alpha p} (F[A_k])^{(d-\alpha p)/d}. \tag{13.8}$$

Recall that we are assuming $0 < \alpha < d/q$ and also $r_c(f) > qd\alpha/(d - q\alpha)$ (the last assumption did not feature in [15], but in fact we do need it). Let $p > q$ be chosen so that $r_c(f) > d\alpha p/(d - \alpha p)$ as well as $\alpha p < d$. Setting $s = \alpha p$ in Lemma 13.3, we get that the expression (13.8) is finite. Thus the first term in the right hand side of (13.7) is bounded by a constant independent of n.

The second term in the right hand side of (13.7) is bounded by

$$\begin{aligned} &C(p)\left(\int_0^1 P\left[\max_{1\le i\le n} |X_i|^{\alpha p} \ge t n^{(d-\alpha p)/d}\right] dt \right. \\ &\quad \left. + \int_1^\infty P\left[\max_{1\le i\le n} |X_i|^{\alpha p} \ge t n^{(d-\alpha p)/d}\right] dt\right) \\ &\le C(p)\left(1 + n\int_1^\infty P[|X_1|^{\alpha p d/(d-\alpha p)} \ge t^{d/(d-\alpha p)} n] dt\right). \end{aligned}$$

By Markov's inequality together with the assumption $r_c(f) > d\alpha p/(d - \alpha p)$, this last integral is bounded by a constant independent of n.

Therefore the expression (13.7) is bounded independently of n, so we can apply Lemma 13.1 to get the L^q convergence in (13.1). □

Proof of Corollary 13.1. Suppose for some $\beta > d$ that $f(x) = \Theta(|x|^{-\beta})$ as $|x| \to \infty$. Then it is easily verified that given $\alpha \in (0, d)$, the condition $I_{1-\alpha/d}(f) < \infty$ implies that $-\beta(1 - \alpha/d) + d < 0$ and hence $\beta > d^2(d - \alpha)^{-1}$. Moreover, it is also easily checked that $r_c(f) = \beta - d$ so that if $\beta > d^2(d - \alpha)^{-1}$ then $r_c(f) > d\alpha/(d - \alpha)$.

Therefore, if $I_{1-\alpha/d}(f) < \infty$ we can apply the case $q = 1$ of Theorem 13.3 to get (13.1) with L^1 convergence. □

The proof of Theorem 13.3 shows that

$$\sup_n \mathbb{E}\left[(n^{1/d} D(X_1, \mathcal{X}_n))^{\varepsilon}\right] < \infty, \tag{13.9}$$

if $\varepsilon > 0$ is such that $\varepsilon d/(d-\varepsilon) < r_c(f)$. The proof of Theorem 13.4, given below, shows that the condition $\varepsilon d/(d-\varepsilon) < r_c(f)$ cannot be dropped in general.

Proof of Theorem 13.4. Let $0 < \alpha < d$. Suppose that $r_c(f) < \alpha d/(d-\alpha)$, and (13.3) holds for some $k_0 \in \mathbb{N}$. Choose r, s such that $r_c(f) < r < s < \alpha d/(d-\alpha)$. Then $M_r(f) = \infty$, so $\sum_k 2^{rk} F(A_k) = \infty$ and therefore there is an infinite subsequence $\mathcal{K}$ of $\mathbb{N}$ such that

$$2^{sk} F(A_k) \geq 1, \quad k \in \mathcal{K}. \tag{13.10}$$

Indeed, if no such $\mathcal{K}$ existed, then for all but finitely many k we would have $2^{rk} F(A_k) \leq 2^{(r-s)k}$ which is summable in k.

Given $k \in \mathbb{N}$, and set $n(k) = \lceil (F(A_k))^{-1} \rceil$, the smallest integer not less than $(F(A_k))^{-1}$. Let E_k be the event that $X_1 \in A_k$ but $X_i \notin A_{k-1} \cup A_k \cup A_{k+1}$ for $2 \leq i \leq n(k)$. Then by the condition (13.3), there is a strictly positive constant c, independent of k, such that for $k \geq k_0$ we have

$$P[E_k] = F(A_k)(1 - F(A_{k-1} \cup A_k \cup A_{k+1}))^{n(k)-1} \geq cF(A_k).$$

If E_k occurs then $D(X_1, \mathcal{X}_{n(k)}) \geq 2^{k-1}$, so for $n = n(k)$ we have (for a different constant c) that

$$\begin{aligned} \mathbb{E}[n^{-1} S_{n,\alpha}] = \mathbb{E}\left[(n^{1/d} D(X_1, \mathcal{X}_n))^{\alpha}\right] &\geq n^{\alpha/d} \mathbb{E}\left[D(X_1, \mathcal{X}_n)^{\alpha} \mathbf{1}_{E_k}\right] \\ &\geq cn^{\alpha/d} F(A_k) 2^{k\alpha} \geq c(F(A_k))^{1-\alpha/d} 2^{k\alpha}. \end{aligned}$$

By (13.10), for $k \in \mathcal{K}$ this lower bound is at least a constant times $2^{k(\alpha - s(d-\alpha)/d)}$, and therefore tends to infinity as $k \to \infty$ through the sequence $\mathcal{K}$, concluding the proof of part (i).

For part (ii), for each $k \geq 2$ choose, in an arbitrary way, a unit radius ball B_k that is contained in A_k. Given $r \in (0, \alpha d/(d-\alpha))$, consider the density function f with $f(x) = C 2^{-rk}$ for $x \in B_k, k \geq 2$, and with $f(x) = 0$ for $x \in \mathbb{R}^d \setminus \cup_{k=2}^{\infty} B_k$; here the normalizing constant C is chosen to make f a probability density function. This gives $F(A_k) = C\omega_d 2^{-rk}$ for each $k \geq 2$; it is easy to see that this f has $r_c(f) = r$, and that (13.3) holds with $k_0 = 3$. Also, for any $\rho > 0$ we have $I_\rho(f) = \omega_d C^\rho \sum_{k \geq 2} 2^{-r\rho k}$ which is finite, so in particular $I_{1-\alpha/d} < \infty$. This choice of f is bounded but not continuous, but can easily be modified to a continuous density with the same properties, e.g., by modifying f in an annulus near the boundary of each ball B_k so as to make it continuous, and then adjusting the normalizing constant C accordingly. □

Acknowledgements Research of Matthew Penrose supported in part by the Alexander von Humboldt Foundation through a Friedrich Wilhelm Bessel Research Award. Research of J.E. Yukich supported in part by NSF grant DMS-0805570.

References

1. Baryshnikov Y, Yukich JE (2005) Gaussian limits for random measures in geometric probability. Ann Appl Probab 15:213–253
2. Beirlant J, Dudewicz E, Györfi L, Meulen E (1997) Non-parametric entropy estimation: An overview. Int J Math Stat Sci 6(1):17–39
3. Bickel P, Breiman L (1983) Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. Ann Probab 11:185–214
4. Costa J, Hero III A (2006) Determining intrinsic dimension and entropy of high-dimensional shape spaces. In: Krim H, Yezzi A (eds) Statistics and analysis of shapes. Birkhäuser, Basel, pp 231–252
5. Evans D, Jones AJ, Schmidt WM (2002) Asymptotic moments of near-neighbour distance distributions. Proc R Soc Lond A Math Phys Sci 458:2839–2849
6. Feller W (1971) An Introduction to probability theory and its applications, vol 2, 2nd edn. Wiley, New York
7. Havrda J, Charvát F (1967) Quantification method of classification processes. Concept of structural α-entropy. Kybernetika (Prague) 3:30–35
8. Jiménez R, Yukich JE (2002) Strong laws for Euclidean graphs with general edge weights. Stat Probab Lett 56:251–259
9. Kozachenko LF, Leonenko NN (1987) A statistical estimate for the entropy of a random vector. Probl Inf Transm 23:95–101
10. Leonenko NN, Pronzato L, Savani V (2008) A class of Rényi information estimators for multidimensional densities. Ann Stat 36:2153–2182
11. Loftsgaarden DO, Quesenberry CP (1965) A nonparametric estimate of a multivariate density function. Ann Math Stat 36:1049–1051
12. Penrose MD (2007a) Gaussian limits for random geometric measures. Electron J Probab 12:989–1035
13. Penrose MD (2007b) Laws of large numbers in stochastic geometry with statistical applications. Bernoulli 13:1124–1150
14. Penrose MD, Yukich JE (2001) Central limit theorems for some graphs in computational geometry. Ann Appl Probab 11:1005–1041
15. Penrose MD, Yukich JE (2003) Weak laws of large numbers in geometric probability. Ann Appl Probab 13:277–303
16. Ranneby B, Jammalamadaka SR, Teterukovskiy A (2005) The maximum spacing estimation for multivariate observations J Stat Plan Inference 129:427–446
17. Rényi A (1961) On measures of information and entropy. In: Proceedings of the 4th Berkeley symposium on mathematics, statistics and probability, vol 1960. University of California Press, Berkeley, pp 547–561
18. Shank N (2009) Nearest-neighbor graphs on the Cantor set. Adv Appl Probab 41:38–62
19. Wade A (2007) Explicit laws of large numbers for random nearest neighbor type graphs. Adv Appl Probab 39:326–342
20. Yukich JE (1998) Probability theory of classical euclidean optimization problems. Lecture notes in mathematics, vol 1675. Springer, Berlin
21. Zhou S, Jammalamadaka SR (1993) Goodness of fit in multidimensions based on nearest neighbour distances. J Nonparametr Stat 2:271–284

Chapter 14
Nonparametric and Probabilistic Classification Using NN-balls with Environmental and Remote Sensing Applications

Bo Ranneby and Jun Yu

Abstract National and international policies today require environmental follow-up systems that detect, in a quality assured way, changes over time in land use and landscape indicators. Questions related to environmental health and spatial patterns call for new statistical tools. We present in this paper some new developments on the classification of land use by using multispectral and multitemporal satellite images, based on techniques of nearest neighbour balls. The probabilistic classifiers introduced are useful for measuring uncertainty at pixel level and obtaining reliable area estimates locally. Also some theoretical considerations for the reference sample plot method (today named k-NN method in natural resource applications) are presented.

14.1 Introduction

Environmental pressures on ecosystems and the continuous loss of suitable habitats for different species jeopardize the ambitions of a sustainable development. Environmental work needs to be enforced at a global, national, regional and local level.

Spatial patterns at different scales are important characteristics of environmental quality. Spatial elements that indicate high environmental quality are usually quite rare and important changes are sometimes very subtle. Existing monitoring programs linked to landscape surveillance do not permit evaluation of data at a local level or of smaller regions. The landscape perspective requires wall-to-wall analysis that are scale-independent, i.e., the results can be used at different levels and the accuracy of the measurements can be clearly defined.

Remote sensing of satellite images offers great potential to assess wall-to-wall changes in the health of ecosystems and identify risks. So far, however, little of this potential has been realized.

Bo Ranneby (✉)
Centre of Biostochastics, Swedish University of Agricultural Sciences, 901 83, Umeå, Sweden
e-mail: bo.ranneby@sekon.slu.se

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_14,

Questions related to environmental health, biodiversity, and spatial patterns in general require new statistical methods. Methods and models must be developed under so called non-standard conditions – conditions under which many statistical methods do not work properly but frequently appear in environmental and remote sensing applications.

In this paper the focus will be on the classification of land use using multispectral and multitemporal satellite images and some convergence results for nearest neighbour balls (NN-balls). Section 14.2 is devoted to classification methods, including a brief review of the conventional classification methods, new developments of non-parametric classification, and the probabilistic classifiers. The reference sample plot method (today named k-NN) is very popular in natural resource applications. Unfortunately there are no theoretical considerations about the selection of a suitable metric and the weights. In Sect. 14.3 we will use NN-balls to show how the weights should be calculated. Some final remarks are given in Sect. 14.4.

14.2 Classification

Supervised classification of remote-sensing images has been widely used as a powerful means to extract various kinds of information concerning the earth environment. The objective of supervised classification in remote sensing is to identify and partition the pixels comprising the noisy image of an area according to its class, with the parameters in the model for pixel values estimated from training samples (ground truths). Usually, the spectral signature is the main aspect of the classes used to classify the pixels. For multispectral image data acquired by remote sensing devices, one should be aware that they are very complex entities which have not only spectral attributes (with correlated bands) but also spatial attributes. Proper utilization of this spatial contextual information, in addition to spectral information, can improve the classification performance significantly in many applications compared to the conventional noncontextual rules such as discriminant analysis and the k-nearest neighbour classifier.

A typical problem with the parametric classification methods is the assumption of (approximative) normality for the feature vector, which is often not fulfilled in remote sensing applications. For non-normal densities it is nevertheless possible to use maximum likelihood (ML) classifiers but with other models for the class distributions. For example, in [18] mixtures of normal distributions are used for classification of images. Their results indicate a substantial increase in correct classification rates compared to classification under normal densities. It is worth noting that mixture models are examples of models when traditional estimation methods such as ML have a tendency to fail. It has been observed that in remotely sensed data feature vectors quite frequently possess bimodal or multimodal empirical distributions. If the training sets are objectively selected empirical distributions of this kind will become even more common. Therefore general estimation methods, which give efficient and robust estimates also when traditional methods break down, are

of fundamental importance for parametric classification. For recent achievements in this direction we refer to [7, 14, 19].

One way to utilize information from neighbouring pixels is to use the wavelet transform and denoising techniques. Yu and Ekström [23] evaluated various traditional classifiers by using the wavelet transform on multispectral images, and found that no matter which type of spatial correlation between pixels the image has and which kind of conventional classification method one uses, the classification by using the wavelet transform (specially the wavelet shrinkage) yields a much lower classification error than that working directly on the original image data, as soon as we have a class image with some structure in it. It is worth noting that this study is based on a model for feature vector assuming normality of class distributions.

Concerning the conventional nonparametric classification methods such as the k-nearest neighbour classifier, its successful combination with the wavelet transform and information theory [15, 25] has shown a great potential in remote sensing applications. The methodology will be described in the following subsections.

We assume that a given pixel from the scene belongs to one of a fixed number of classes, say, $C_1, \ldots, C_K$. The proportion of pixels belonging to class C_j in the population under study is denoted by π_j, which is usually unknown. Each pixel gives rise to certain measurements (e.g., spectral signatures in remote sensing images, or their various transformations), which form the d-dimensional feature vector $\mathbf{X} \in \mathbb{R}^d$. Our task is to allocate each pixel into one of the classes in $\{C_1, \ldots, C_K\}$, on the basis of the observed value $\mathbf{X} = \mathbf{x}$ and/or prior information about the classes.

14.2.1 Conventional Classification Methods

Here follows a short description of three traditional classification methods. See [11, 16] for a comprehensive review of these methods.

14.2.1.1 Parametric Model

In parametric models, the feature vectors from class C_j are assumed to be distributed according to the density $p_j(\mathbf{x})$. Then the posterior distribution of the classes after observing $\mathbf{x}$ is $p(C_j|\mathbf{x}) \propto \pi_j p_j(\mathbf{x})$.

If we assume the probability model in which the observations for class C_j are (d-dimensional) multivariate normal with mean $\boldsymbol{\mu}_j$ and covariance matrix Σ_j, the Bayes rule is to allocate a future observation $\mathbf{x}$ to the class that minimises

$$Q_j = -2 \log p(C_j|\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_j)^T \Sigma_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) + \log |\Sigma_j| - 2 \log \pi_j .$$

This method is known as *quadratic discriminant analysis* (QDA). When the classes have a common covariance matrix Σ, minimizing Q_j is equivalent to maximizing the linear terms

$$L_j = 2\mathbf{x}^T \Sigma^{-1} \boldsymbol{\mu}_j - \boldsymbol{\mu}_c^T \Sigma^{-1} \boldsymbol{\mu}_j + 2 \log \pi_j,$$

which leads to the *linear discriminant analysis* (LDA).

In practice, we replace $\boldsymbol{\mu}_j$, Σ_j or Σ, and π_j with their estimates (e.g., maximum likelihood) by using a training sample.

14.2.1.2 Nonparametric Model

In nonparametric designs we do not assume any model of the probability distributions. There is a number of nonparametric classifiers based on nonparametric estimates of the class densities or of the log posterior, such as kernel methods, orthogonal expansion, projection pursuit, and others. Here we use the simple adaptive kernel method which gives a classifier known as the *k-nearest neighbour rule*. The k-NN rule is introduced by [6], which is one of the most theoretically elegant and yet simple classification techniques. It is based on finding the k nearest (usually in the Euclidean distance) pixels from the training samples (prototypes), and taking a majority vote amongst the classes of these k samples, or equivalently, estimating the posterior distributions $p(C_j|\mathbf{x})$ by the proportions of the classes amongst the k samples. The nearest neighbour rule (when k is 1) is of special interest in our later applications.

Nonparametric classification is often associated with the notion "prototype". A prototype is a representative point of certain class in the feature space $\mathbb{R}^d$ and is attached with both field and satellite information. Different classes can have different number of prototypes, depending on the property of field data or ground truth we have. Prototype selection plays a key role in the nonparametric classification. k-NN is a prototype classifier.

Specifically, assuming the number of prototypes for class C_j is n_j and $n = \sum_{j=1}^{K} n_j$ is the number of all prototypes. For an arbitrary pixel with observed feature value $\mathbf{x}$, let $B(\mathbf{x})$ denote the hypersphere with centre $\mathbf{x}$ and containing exactly k prototypes and $||B(\mathbf{x})||$ the volume of $B(\mathbf{x})$. The number of prototypes in $B(\mathbf{x})$ from class C_j is denoted by k_j. The unconditional probability density of $\mathbf{x}$ and the class-conditional probability density for class C_j can be approximated as

$$p(\mathbf{x}) \approx \frac{k}{n||B(\mathbf{x})||} \quad \text{and} \quad p_j(\mathbf{x}) \approx \frac{k_j}{n_j||B(\mathbf{x})||}, \tag{14.1}$$

respectively. Then the posterior probabilities are obtained as

$$p(C_j|\mathbf{x}) = \frac{p_j(\mathbf{x})\pi_j}{p(\mathbf{x})} \approx \frac{\frac{k_j}{n_j||B(\mathbf{x})||} \cdot \frac{n_j}{n}}{\frac{k}{n||B(\mathbf{x})||}} = \frac{k_j}{k}, \tag{14.2}$$

which implies that the k-NN classification rule assigns the class label independent of the volume $||B(\mathbf{x})||$.

When we are dealing with sparse or rarely occurring objects, however, it is unrealistic to assume that the number of prototypes for these classes is large so the approximation above may be violated. Thus there is an obvious risk that such classes will be underestimated. To overcome this problem alternatives of deriving the probabilities are needed.

The k-NN classifier uses the whole prototype set of any class to choose the neighborhood, which differs from using the k-nearest neighbour density estimate for each class. Ties in distances can occur with finite-precision data. One solution to this problem is to include in the vote all distances equal to the kth largest.

Notice that "nearest" implies a certain metric over the feature space $\mathbb{R}^d$. No matter which metric is used, the k-NN rule is applied in the same way. However, this metric has to be defined properly because the closeness in the feature space is no longer so simple as in Euclidean case. This is connected to the similarity in spatial statistics, which means that the smaller the distance, the higher the similarity between input data and the prototype.

14.2.2 Nonparametric Classification: New Developments

Very often objects with high environmental values (such as old deciduous forest) are rare. Remote sensing classification of sparse and rarely occurring objects in the landscape is complicated and reported classification rates are usually poor. For several reasons traditional methods will not give satisfactory results:

- The number of field plots in different classes varies too much leading to an unfavorable treatment of sparsely occurring objects.
- Data editing methods, like the multiedit algorithm [16], often drop whole classes when applied to moderately sized training sets with more dimensions and classes.
- Noncontextual methods neglect information from neighbouring plots and the feature vector is not normally distributed.
- Parametric contextual classification methods do utilize information from neighbouring plots but the demands for field data increases, and estimation of some parameters in the model is unreliable and in many situations impossible.
- The feature vector is not normally distributed and the different class distributions are often highly overlapping.

To overcome these problems nonparametric approaches have been used. Information from neighbouring pixels will be utilized by using the wavelet transform to denoise the original images by removing small wavelet coefficients and inverting the wavelet transform. Hereafter a suitable classification algorithm is applied. For previous work in this direction see [3, 15, 20, 22–26].

The nonparametric classification that we have developed is based on a new approach for classification of multitemporal satellite data sets, combining multispectral and change detection techniques. The classification method has to be adapted for handling a mixture of reference data from different sources in order to compensate

Table 14.1 Definition of forest classes

Proportion deciduous		Age			
		10–39	40–69	70–109	110+
	Pine $\geq$ 70%	1	7	13	19
0–19%	Spruce $\geq$ 70%	2	8	14	20
	"Remaining"	3	9	15	21
20–49%		4	10	16	22
50–69%		5	11	17	23
$\geq$70%		6	12	18	24
Class 25 = basal area $\leq 1\ m^2\ ha^{-1}$					

for the lack of ground samples from rare classes within the main reference data source.

As an example, classification of forest land in a study area located in the center of Sweden is presented. In this application the forest is defined into 24 different classes with respect to species composition and age, supplemented with the class "low basal area", when species composition and age do not have any meaning. Table 14.1 presents the definition of those 25 forest classes to be used.

The field data from the Swedish National Forest Inventory (NFI) and satellite images with different dates were used. The dimension of the feature space is 12.

Our classification algorithm is based on the nearest neighbour method. The main procedures are as follows.

- Define the target function. The probabilities of correct classification for each class are used, so that each class is of equal importance. Note that it differs from the overall correct classification, which is often used in various applications.
- Denoise the feature vector. The wavelet shrinkage method based on 2D-wavelet transform is used to denoise the images. The feature vector consists of components that are pixel values from different spectral bands of different temporal images over both years and seasons.
- Remove outliers from the reference data. When the feature vector has high dimension and the number of classes is large, new data editing methods are needed in order to remove outliers due to poor quality of field data, and to find out the prototypes for each class in the case of nearest neighbour classifiers. This must be done so that poor quality data is removed at the same time as most of the natural variations within the different classes are kept. This is achieved by working with a further subdivision of the classes with respect to other information, as e.g. basal area, and for the new classes (now 109 classes) suitable techniques for outlier removal are applied.
- Calculate the information values in the components in the feature vector. Before the feature vector can be used the components have to be rescaled. The size of the rescaling factor is a function of the dependence between the studied objects and the component. Since the classes are defined by more than one variable, there will not exist any natural order relation between the classes so conventional dependence measures, such as the correlation coefficient, cannot be used.

Instead measures with the origin in information theory are used. Here we apply the normed information rate, $IR_{\xi,\eta}$, introduced by [13], which measures the divergence between the general 2-dimensional distribution and the 2-dimensional distribution calculated under the assumption about independence and then uses Shannon's entropy for the general 2-dimensional distribution to get a measure between zero (independence) and one (perfect dependence). The definitions are as follows:

$$IR_{\xi,\eta} = \frac{I_{\xi,\eta}}{H_{\xi,\eta}}, \tag{14.3}$$

where

$$I_{\xi,\eta} = \sum_{i,j} P(\xi = i, \eta = j) \log \frac{P(\xi = i, \eta = j)}{P(\xi = i)P(\eta = j)}$$

is the Kullback–Leibler information rate for the class variable ξ and the feature component variable η, and

$$H_{\xi,\eta} = -\sum_{i,j} P(\xi = i, \eta = j) \log P(\xi = i, \eta = j)$$

is the Shannon joint entropy for (ξ, η).

- Determine a proper metric. We prefer to use the following metric:

$$\sqrt{\sum_{i=1}^{d} w_i^2 \left| \frac{x_i(\boldsymbol{s}) - x_i(\mathbf{t})}{q_i} \right|^2}. \tag{14.4}$$

Here $\mathbf{x}(\boldsymbol{s}) = [x_1(\boldsymbol{s}), x_2(\boldsymbol{s}), \ldots, x_d(\boldsymbol{s})]$, $\mathbf{x}(\boldsymbol{t}) = [x_1(\boldsymbol{t}), x_2(\boldsymbol{t}), \ldots, x_d(\boldsymbol{t})]$ are two pixels at location $\boldsymbol{s}$ and $\boldsymbol{t}$, with attribute of the spectral components, d is the dimension of feature vector, q_i is the inter-quartile range of the ith feature component, and the weight

$$w_i = [IR_{\text{class, feature component}}(i)]^p \tag{14.5}$$

is determined by the normed information rate of the ith feature component obtained from (14.3), raised to some power p. And p is determined so the target function is maximized. Usually p equals to 1.

- Determine prototypes for the classes. Here the information about the occurrence of different classes in the reference data is used.
- Run a nonparametric classification. Here the nearest neighbour classifier is used.
- Declare the quality of classification result by using probability matrices. The probability matrix is based on the confusion matrix and defined as $\mathbf{P} = [P_{ij}]$ where P_{ij} is the probability that class C_i is classified as class C_j.

As a result in the example, the probability for correct classification varies between 30 and 40% for most of the pure classes although it is considerably lower for some other classes, especially those with different types of mixed forest. The

Table 14.2 Probabilities of correct classification for C_1 and C_{24}

Class	LDA with prior	LDA without prior	Our method
C_1	70%	12%	39%
C_{24}	0%	34%	37%

largest class is C_1 (younger pine forest) and the smallest is C_{24} (older deciduous forest). We have compared our classification rates with LDA with or without prior information. Note that it was impossible to apply QDA here because the field plots for small classes were too few to be able to estimate the covariances. LDA with prior information is the most commonly used method. With this method the probability for correct classification is 0% for 14 out of the 25 classes. The reason is that all small classes essentially are classified as the larger classes, particularly C_1, which gets a rate of 70% for correct classification. If we do not use any prior information then the rates increases for the small classes but drops dramatically (to 12%) for class C_1. A comparison for the largest and smallest classes is given in Table 14.2.

As we want to have high classification accuracy also for small classes they have to be overestimated (unless the classification is 100% perfect). To overcome this problem we can calculate matrices for area correction and the precision in the correction matrices, using Monte Carlo techniques. From this it is possible to obtain unbiased estimates for the different classes and standard errors of the area estimates. Resulting estimated areas and standard errors are omitted here.

It is worth to notice that the estimation property is reliable only on the scene level and it is inappropriate to apply it to the local area or pixel level. To obtain area estimation of good quality (such as unbiasedness), a new approach is needed, and we will take it up in the coming subsection.

14.2.3 Probabilistic Classifier

For several applications and a large number of users it is necessary to give pixelwise information. However, as the distributions for different classes are highly overlapping it is not possible to get satisfactory accuracy at pixel level. It is, therefore, necessary to introduce a new concept, pixelwise probabilistic classifiers. Instead of classifying each pixel to a specific class, each pixel is given a probability distribution describing how likely the different classes are. This new concept has several important applications: (a) it is possible to give quality declarations at pixel level, (b) it is possible to derive unbiased area estimates also for small areas and deviating areas, (c) it is possible to derive cost-efficient monitoring systems.

The concept of pixelwise probabilistic classifiers is closely related to fuzzy classification; see [2] for different definitions. The drawback with fuzzy classifiers is that until now the methods are not based on any rigorous theory and the literature is sparse where the probabilistic classifier definition is used.

In principle, probabilistic classifiers can be derived both from parametric and nonparametric classification methods. However, deriving probabilistic classifiers from parametric methods is a complicated task since for several classes the distributional assumptions are not satisfied. Furthermore, it is of key importance to have consistent parameter estimates. To avoid the problem of misspecification, nonparametric approaches are preferred. The probabilistic classifier derived here is based on the k-NN method with $k = 1$, defined in Sect. 14.2.1.2.

Articles on fuzzy classification and k-NN are available, see e.g. [1, 10], but they have a different scope. To be useful for area estimation and quality assessment at pixel level the probabilities must be extracted from proper probability distributions for the different classes. When the classifier uses distances in feature space to derive the probabilities it is intuitively obvious that the probabilities will be inversely proportional to the distances raised to some power. However, there is only one value that is correct and that value depends on the dimension of the feature vector. Actually, by using the nonparametric density estimation [12], the density for class C_j at $\mathbf{x}$ will be inversely proportional to the volume of the nearest neighbour ball with center $\mathbf{x}$. Note that the volume is proportional to the radius, based on the metric defined in (14.4), raised to the dimension of the feature space. The derivation is similar as for (14.1) and (14.2), but now using the NN-balls in the feature space.

As a theoretical background, we introduce a proposition based on [14] before the derivation. Let $B(\mathbf{x}, r) = \{\mathbf{y} : |\mathbf{x} - \mathbf{y}| \leq r\}$ denote the ball of radius r with center at $\mathbf{x}$ in the feature space. Then the volume of the ball $B(\mathbf{x}, r)$ is $\|B(\mathbf{x}, r)\| = c_d r^d$ where $c_d = \pi^{d/2}/\Gamma(d/2 + 1)$.

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a sequence of i.i.d. d-dimensional random vectors with the common density function $g(\mathbf{x})$ and define the nearest neighbour distance to the point $\mathbf{X}_i$ by

$$R_n(i) = \min_{j \neq i} |\mathbf{X}_i - \mathbf{X}_j|.$$

Then $B(\mathbf{X}_i, R_n(i))$ is the so-called NN-ball of $\mathbf{X}_i$.

Proposition 14.1. *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be i.i.d. random vectors with the common density function $g(\mathbf{x})$ and*

$$Y_n(i) = n \, \|B(\mathbf{X}_i, R_n(i))\| = n c_d R_n^d(i).$$

Then

$$(\mathbf{X}_i, Y_n(i)) \xrightarrow{D} (\mathbf{X}, Y), \quad \text{as } n \to \infty, \tag{14.6}$$

where $\mathbf{X}$ has density $g(\mathbf{x})$ and Y given $\mathbf{X} = \mathbf{x}$ has an exponential distribution with parameter $g(\mathbf{x})$.

Note that $(\mathbf{X}_i, Y_n(i)), i = 1, 2, \ldots, n$ are exchangeable and asymptotically independent, and the joint probability density of $(\mathbf{X}, Y)$ is $g^2(\mathbf{x}) \exp\{-y g(\mathbf{x})\}$.

Furthermore,

$$E[Y_n(i)|\mathbf{X}_i = \mathbf{x}] \to E[Y|\mathbf{X} = \mathbf{x}] = 1/g(\mathbf{x}).$$

Let now

$$R_n(\mathbf{x}, C_j) = \text{the minimum distance in the feature space (using the metric (14.4)) from } \mathbf{x} \text{ to the prototypes of class } C_j. \tag{14.7}$$

The conditional probability density for class C_j can be estimated by

$$\hat{p}_j(\mathbf{x}) = \frac{1}{n_j \left\| B(\mathbf{x}, R_n(\mathbf{x}, C_j)) \right\|}. \tag{14.8}$$

Then the posterior probabilities are obtained as

$$\begin{aligned}
\hat{p}(C_j|\mathbf{x}) &= \frac{\pi_j p_j(\mathbf{x})}{\sum_{i=1}^{K} \pi_i p_i(\mathbf{x})} \\
&= \frac{\pi_j}{n_j \left\| B(\mathbf{x}, R_n(\mathbf{x}, C_j)) \right\|} \Bigg/ \sum_{i=1}^{K} \frac{\pi_i}{n_i \left\| B(\mathbf{x}, R_n(\mathbf{x}, C_i)) \right\|} \\
&\overset{\pi_j \simeq n_j/n}{\simeq} \frac{1}{\left\| B(\mathbf{x}, R_n(\mathbf{x}, C_j)) \right\|} \Bigg/ \sum_{i=1}^{K} \frac{1}{\left\| B(\mathbf{x}, R_n(\mathbf{x}, C_i)) \right\|} \\
&= \frac{R_n^{-d}(\mathbf{x}, C_j)}{\sum_{i=1}^{K} R_n^{-d}(\mathbf{x}, C_i)}.
\end{aligned}$$

For each pixel with location $\boldsymbol{s}$ and feature vector $\mathbf{x}(\boldsymbol{s}) = [x_1(\boldsymbol{s}), x_2(\boldsymbol{s}), \ldots, x_d(\boldsymbol{s})]$, we assign the following proper probability to class C_j

$$p^j(\boldsymbol{s}) = \frac{R_n^{-d}(\mathbf{x}(\boldsymbol{s}), C_j)}{\sum_{i=1}^{K} R_n^{-d}(\mathbf{x}(\boldsymbol{s}), C_i)}. \tag{14.9}$$

Thus, the probability vector at pixel $\boldsymbol{s}$ is obtained as $\mathbf{p}(\boldsymbol{s}) = [p^1(\boldsymbol{s}), p^2(\boldsymbol{s}), \ldots, p^K(\boldsymbol{s})]$.

14.2.4 Applications of the Probabilistic Classifier

The probabilistic classifiers can be used in different ways. For instance, it can be used to measure uncertainty at pixel level, or to obtain reliable area estimates locally. It can also be applied to correct the erroneous mask of maps (see comments in the last section).

- *Quality evaluation of traditional classification*
  A traditional classification is obtained by choosing the class with the highest probability for each pixel. With the probabilistic classifier, it is possible to give

quality declaration at pixel level. The pixelwise vectors of probabilities (14.9) can be used to judge how reliable a traditional classification is and to derive measures of the uncertainty (entropy) for the individual pixels. Based on the probability vector, the entropy at pixel $\boldsymbol{s}$ is defined as

$$-\sum_{j=1}^{K} p^j(\boldsymbol{s}) \log p^j(\boldsymbol{s}). \tag{14.10}$$

- *Monitoring systems*
  In application, in addition to the classified map a raster map expressing the quality (entropy defined in (14.10)) for each pixel is produced. Areas where the precision is unsatisfactory will be identified. In these areas it is possible to take additional field plots and perform an improved classification. Here it is important that the classification method can handle different types of reference data. The final product will be two maps, one showing the pixel wise classification and the other giving the classification accuracy at pixel level. Figure 14.1 illustrates our concept of probabilistic classifiers operated on a small site of size 16×16 pixels. The area consists of mixed forest, which implies difficulty to classify.

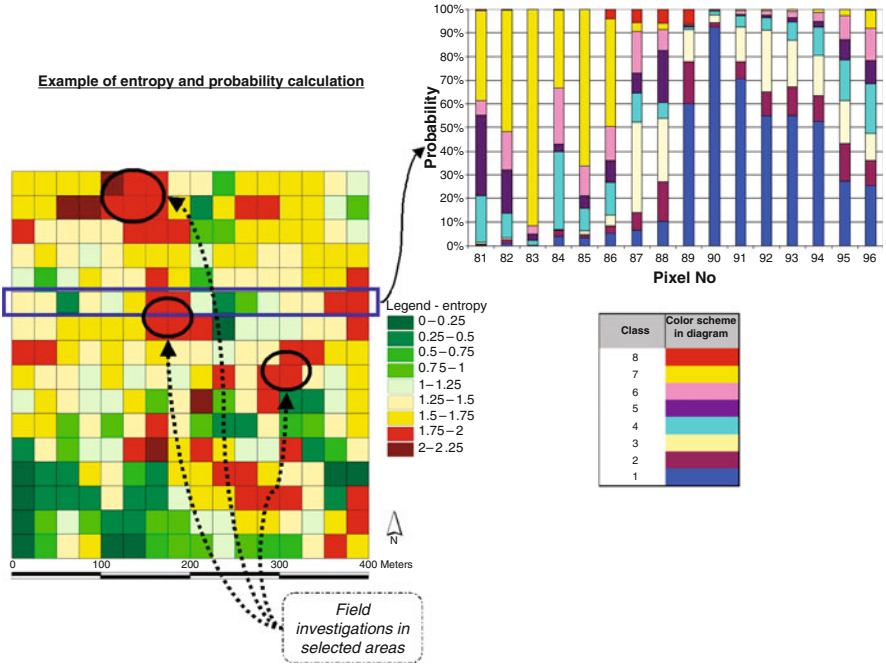


Fig. 14.1 The concept of probabilistic classifiers. Test area with mixed forest, eight classes. Field plots can be added in areas where entropy is high (i.e., quality at pixel level is unsatisfactory) to produce additional reference data as input to a renewed (improved) classification

Darkest colored pixels indicate higher entropy values or unsatisfactory quality. The diagram on top right shows the probability distributions of 16 pixels (one row) picked from the entropy map on the left side. For a given pixel, the more even distributed probabilities among classes, the high entropy. If one class dominates the probability vector (i.e., having high probability), then the pixel has very low entropy and is reliably classified.

- *Area estimation*
  Although it is possible to obtain satisfactory area estimates on scene level using area correction matrices derived from the confusion matrices as we discussed in Sect. 14.2.2, the precision is not good enough for smaller areas or on pixel level. The pixelwise vectors of probabilities can be used to get unbiased area estimates also for quite small areas and for deviating areas, such as potential areas for nature reserves. The area estimates is obtained simply by summation of the "class probabilities" over the selected subregion. It is extremely important that proper probability distributions allowing frequency interpretation are derived; otherwise misleading results are obtained.

  Concerning the quality assessment related to our new area estimates, resampling and subsampling methods will be considered. The advantage of such methods is that statistical inference can be done without knowledge of the underlying spatial dependence mechanism and marginal distributions that generated the data. Furthermore, for the user no explicit theoretical derivation is necessary but instead intensive computing. Most of the proposed methods in the literature assume stationary data. However, when modeling real-life data, the hypothesis of stationarity often must be rejected. Especially for nonstationary spatial data, there are very few available methods to do inference. Sherman [17] has provided a short discussion on subsampling of spatial lattice data in the context of regression models with stationary errors, and later [4, 5] has proposed resampling and subsampling methods for nonstationary spatial lattice data. The nonparametric variance estimator proposed by them can be used to estimate not only the variance of the area estimates derived from the probabilistic classification, but also the variance of estimators for continuous variables (such as wood volume) with respect to (estimated) area of interest (e.g., forest).

14.3 *k*-NN Estimation in Natural Resources

Starting around 1990 with articles by [9, 21], k-NN techniques have gained popularity for natural resources applications. Kilkki and Pivinen [9] used the term "reference sample plot method". Later the term "k-NN method" was established although the method differs from the k-NN method introduced by [6]. Katila and Tomppo [8] provided a survey of the early k-NN (reference sample plot method) literature for estimating forest attributes. For prediction we have observations of the response variable Y for a subset of pixels and observations of covariates $\mathbf{X}$ (d-dimensional feature vector) for all pixels. The k-NN prediction for the ith pixel is

$$\tilde{y}_i = \left(\sum_{j=1}^{k} w_{ij} \right)^{-1} \sum_{j=1}^{k} w_{ij} y_{ij},$$

where $\{y_{ij}, j = 1, 2, \ldots, k\}$ is the set of observations closest to the ith pixel with respect to a metric in the feature space, and w_{ij} is the weight assigned to the jth nearest neighbour observation. Common selections for the metric are Mahalanobis distance or Euclidean distance. Usually the weights are chosen to be inversely proportional to the distance raised to the power of 1 or 2, regardless of the dimension of the feature space.

Theoretical considerations about the selection of the weights and a suitable metric are lacking. For selection of a suitable metric some of the ideas described in Sect. 14.2.2 will be useful. In this section we will use NN-balls to show how the weights should be calculated.

For the reference set (e.g., satellite data) we have information both for the response variable Y and the feature vector $\mathbf{X}$. For a pixel with value $\mathbf{x}$, the best predictor of Y, given squared loss function, is the conditional expectation $E[Y|\mathbf{X} = \mathbf{x}]$. It can be written as

$$E[Y|\mathbf{X} = \mathbf{x}] = \int y p(y|\mathbf{x}) dy = \int y \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})} dy. \tag{14.11}$$

The support of Y is partitioned into sets C_j, $j = 1, \ldots, K$. Then

$$\begin{aligned} E[Y|\mathbf{X} = \mathbf{x}] &= \sum_{j=1}^{K} \frac{1}{p(\mathbf{x})} \int_{C_j} y p(\mathbf{x}|y) p(y) dy \\ &= \sum_{j=1}^{K} \frac{1}{p(\mathbf{x})} P(Y \in C_j) \int_{C_j} \frac{y p(y)}{P(Y \in C_j)} p(\mathbf{x}|y) dy. \end{aligned}$$

If the partition $\{\bigcup C_j\}$ is fine enough then $p(\mathbf{x}|y)$ is almost constant for all $y \in C_j$, and denoted by $p(\mathbf{x}|C_j)$. This gives us

$$\begin{aligned} E[Y|\mathbf{X} = \mathbf{x}] &\simeq \sum_{j=1}^{K} \frac{1}{p(\mathbf{x})} P(Y \in C_j) p(\mathbf{x}|C_j) \int_{C_j} \frac{y p(y)}{P(Y \in C_j)} dy \\ &= \sum_{j=1}^{K} \frac{1}{p(\mathbf{x})} P(Y \in C_j) p(\mathbf{x}|C_j) \mu_j. \end{aligned} \tag{14.12}$$

Here μ_j is the expected value of the response variable for class C_j and $p(\mathbf{x}|C_j)$ is the density of the covariates for class C_j.

Suppose now that for the class C_j we have n_j reference points and let $n = \sum_{j=1}^{K} n_j$, the total number of reference points. Suppose that a suitable metric in the feature space, such as (14.4), has been chosen. With respect to that metric the

distance from the pixel with covariate value $\mathbf{x}$ to the nearest pixel in the reference set belonging to class C_j is denoted, as in (14.7), by $R_n(\mathbf{x}, C_j)$. As we have seen before the volume of the nearest ball converges in distribution to an exponential distribution,

$$n_j \left\| B(\mathbf{x}, R_n(\mathbf{x}, C_j)) \right\| = n_j c_d R_n^d(\mathbf{x}, C_j) \xrightarrow{D} Z_j,$$

where Z_j is exponential distributed with mean $1/p(\mathbf{x}|C_j)$. Furthermore,

$$E[n_j c_d R_n^d(\mathbf{x}, C_j)] \to E[Z_j] = \frac{1}{p(\mathbf{x}|C_j)}.$$

To obtain a predictor $\hat{y}(\mathbf{x})$ of the pixel with covariate value $\mathbf{x}$ we substitute each component in the best predictor $E[Y|\mathbf{X} = \mathbf{x}]$ (14.12) with a reasonable estimate. Thus,

$$\hat{y}(\mathbf{x}) = \sum_{j=1}^{K} \frac{1}{\hat{p}(\mathbf{x})} \hat{P}(Y \in C_j) \hat{p}(\mathbf{x}|C_j) \hat{\mu}_j. \tag{14.13}$$

If the reference set is chosen properly, then

$$\frac{n_j}{n} \to P(Y \in C_j).$$

The mean μ_j has to be estimated by the observed value y_j and the density $p(\mathbf{x}|C_j)$ will be estimated by $[n_j c_d R_n^d(\mathbf{x}, C_j)]^{-1}$, and $p(\mathbf{x})$ by $\sum_{j=1}^{K} \frac{n_j}{n} [n_j c_d R_n^d(\mathbf{x}, C_j)]^{-1}$. Thus

$$\begin{aligned} \hat{y}(\mathbf{x}) &= \frac{1}{\hat{p}(\mathbf{x})} \sum_{j=1}^{K} y_j \frac{n_j}{n} [n_j c_d R_n^d(\mathbf{x}, C_j)]^{-1} \\ &= \left(\sum_{j=1}^{K} R_n^{-d}(\mathbf{x}, C_j) \right)^{-1} \sum_{j=1}^{K} y_j R_n^{-d}(\mathbf{x}, C_j). \end{aligned} \tag{14.14}$$

Consequently, the weights in k-NN estimation shall be inversely proportional to the distance raised to the power of d, the dimension of the covariate (feature) space, and not to 1 or 2 which are standard choices in the k-NN (reference sample plot method) literature.

14.4 Final Remarks

Remote sensing classification of sparse and rarely occurring objects in the landscape is complicated. An approach, based on nonparametric statistical methods, has been developed to overcome the problems arisen with traditional classification

methods. It utilises information from multitemporal and multispectral satellite data and change detection techniques. The classification method is also adapted for handling a mixture of reference data from different sources in order to compensate for the lack of ground samples from rare classes within the main reference data source. The new algorithm is derived in order to optimize the average probability for correct classification, i.e., each class is equally important.

As we have seen the probabilistic classifiers are useful for measuring uncertainty at pixel level and obtaining reliable area estimates locally. It can also be applied to correct the masks of maps. The masks from the topographic maps are usually used to define different land use of vegetation, e.g., forest land, peat land, or agricultural land. Then the classification is performed within the mask. It has, however, been found that the errors in masks can be significant, and thus development of better masks is needed. Otherwise we know a priori that the classification accuracy will not be satisfactory. An application of the probabilistic classifiers to obtain new masks for forest land and peatland has been conducted successfully in other studies of the authors. The idea is as follows: (a) use the probabilistic classifier together with information from digital elevation model (wetness index) and neighbouring pixels to get the probability vector for classes (both forest and peatland) at pixel level and consider it as a prior probability distribution for each pixel; (b) apply the probabilistic classifier by using remotely sensed data to get the probability vector for classes at pixel level; (c) join (a) with (b) to get the a posterior probability vector for each pixel; (d) aggregate the classes to two categories: Forest and Peatland; (e) calculate entropies for each pixel and determine the threshold value according to the entropies so that pixels belong to one of the following zones: Forest, Peatland, or Transition Zone; (f) recalculate the probability vectors for pixels in Forest and Peatland zones. In such a way, two masks with higher quality are obtained for forest and peatland, respectively. A third one, the Transition Zone, is given for pixels with a certain degree of uncertainty.

References

1. Bereau M, Dubuisson B (1991) A fuzzy extended k-nearest neighbor rule. Fuzzy Sets Sys 44:17–32
2. Bezdek J, Keller J, Krishnapuram R, Pal N (1999) Fuzzy models and algorithms for pattern recognition and image processing. Kluwer, Norwell
3. Choi H, Baraniuk R (1999) Image segmentation using wavelet-domain classification. SPIE Proceedings, Denver, July 1999, vol 3816, pp 306–320
4. Ekström M (2008) Subsampling variance estimation for nonstationary spatial lattice data. Scand J Stat 35:38–63
5. Ekström M, Sjöstedt-de Luna S (2004) Subsampling methods to estimate the variance of sample means based on nonstationary spatial data with varying expected values. J Am Stat Assoc 99:82–95
6. Fix E, Hodges J (1951) Discriminatory analysis – nonparametric discrimination: consistency properties. Report 4, US Air Force School of Aviation Medicine, Random Field, Texas
7. Jimenez R, Yukich J (2002) Asymptotics for statistical distances based on voronoi tessellation. J Theor Prob 15:503–541

8. Katila M, Tomppo E (2002) Stratification by ancillary data in multisource forest inventories employing k nearest neighbour estimation. Can J For Res 3:1548–1561
9. Kilkki P, Päivinen R (1987) Reference sample plots to combine field measurements and satellite data in forest inventory. In: Remote sensing-aided forest inventory, pp 209–212. Seminars organised by SNS and Taksaattoriklubi, Hyytil, Finland
10. Kissiov V, Hadjitodorov S (1992) A fuzzy version of the k-nn method. Fuzzy Sets Sys 49: 323–329
11. Kuncheva L (2000) Fuzzy classifier design. In: Studies in fuzzness and soft computing, vol 49. Springer, Heidelberg
12. Loftsgaarden D, Quesenberry C (1965) A nonparametric estimate of a multivariate density function. Ann Math Stat 36:1049–1051
13. Rajski C (1964) On the normed information rate of discrete random variables. Translation of the third Prague conference, pp 583–585
14. Ranneby B, Jammalamadaka S, Teterukovskiy A (2005) The maximum spacing estimate for multivariate observations. J Stat Plan Inference 129:427–446
15. Ranneby B, Yu J (2003) Remote sensing classification of environmental indicators. In: Proceedings of 30th international symposium on remote sensing of environment – Information for risk management and sustainable development, Hawaii
16. Ripley B (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge
17. Sherman M (1996) Variance estimation for statistics computed from spatial lattice data. J R Stat Soc Series B 58:509–523
18. Taxt T, Lid Hjort N, Eikvil L (1991) Statistical classification using a linear mixture of two multinormal probability densities. Pattern Recognit Lett 12:731–737
19. Teterukovsky A, Ranneby B (2003) The maximum spacing estimation based on dirichlet tessellation. PhD thesis, Swedish University of Agricultural Sciences. In Doctoral thesis Silvestria 277, Swedish University of Agricultural Sciences
20. Teterukovsky A, Yu J (2002) Contextual reclassification of multispectral images: A markov random field approach. Inf Processes 2:12–21
21. Tomppo E (1991) Satellite imagery-based national forest inventory of Finland. Int Arch Photogramm Remote Sens 28:419–424
22. Yamaguchi Y, Nagai T, Yamada H (1995) JERS-1 SAR image analysis by wavelet transform. IEICE Trans Commun E78-B(12):1617–1621
23. Yu J, Ekström M (2003) Multispectral image classification using wavelets: A simulation study. Pattern Recognit 36(4):889–898
24. Yu J, Ekström M, Nilsson M (2000) Image classification using wavelets with application to forestry. In: Proceedings of the 4th international conference on methodological issues in official statistics, Stockholm
25. Yu J, Ranneby B (2007) Nonparametric and probabilistic classification of agricultural crops using multitemporal images. J Remote Sens 11(5):748–755
26. Yu J, Ranneby B (2007) Nonparametric classification and probabilistic classifiers with environmental and remote sensing applications. In: Lai T, Shao Q, Qian L (eds) Asymptotic theory in probability and statistics with applications, Advances lectures in mathematics, vol 2. International Press, Cambridge

Chapter 15
Probabilistic Recurrence Relations

M. Bhaskara Rao, S. Kasala, and H. Zhang

Abstract A sampling of discrete probability problems, some of them coming from consulting work, is presented. We demonstrate how a probabilistic recurrence relation arises from the pit of the problem and present ways and means of solving the recurrence relation.

15.1 Introduction

Recurrence relations crop up frequently in mathematical research. The basic set-up is as follows. A sequence $X_n, n \geq 0$ of numbers is said to satisfy a recurrence relation of order $k \geq 1$ if

$$X_n = a_1 X_{n-1} + a_2 X_{n-2} + \cdots + a_k X_{n-k} + b_n, n \geq k$$

for some fixed constants $a_1, a_2, \cdots, a_k$, and sequence $b_n, n \geq k$. The goal is to find an explicit mathematical expression for the nth term X_n. In order to find an expression, we need the values of $X_0, X_1, \cdots, X_{k-1}$. These values are collectively called 'boundary conditions'. The mathematical form, of course, depends on the 'boundary conditions'. If it is not possible to find an explicit mathematical expression for the nth term of the sequence, one could investigate the asymptotic behavior of the sequence. If the sequence $b_n, n \geq k$ is a constant, i.e., $b_n = b$ for all $n \geq k$, the recurrence relation is said to be homogeneous. If the relation is homogeneous, one could use 'difference equations' methodology to find an explicit expression for the nth term of the sequence. If the relation is not homogeneous, one has to devise some 'ad hoc' method to solve the recurrence relation.

M. Bhaskara Rao (✉)
Department of Environmental Health, Center for Genome Information, University of Cincinnati, Cincinnati, OH 45267, USA
e-mail: raomb@ucmail.uc.edu

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_15,

A classical example of a recurrence relation is the one satisfied by the Fibonacci sequence $X_n, n \geq 0$, i.e., $X_n = X_{n-1} + X_{n-2}, n \geq 2$ with boundary conditions $X_0 = 0$ and $X_1 = 1$. This is a homogeneous recurrence relation of order 2. The nth term has an explicit expression:

$$X_n = \frac{\phi^n - (1-\phi)^n}{\sqrt{5}}, \quad n \geq 0,$$

where $\phi = \frac{\sqrt{5}+1}{2}$ is the so-called Golden Ratio.

In this article, we focus on recurrence relations that arise from probabilistic investigations. Probabilistic recurrence relations are of two kinds. The sequence $X_n, n \geq 0$ could be a sequence of probabilities satisfying a recurrence relation of order k for some $k \geq 1$. Solving such recurrence relations falls into the realm of mathematics. On the other hand, the sequence $X_n, n \geq 0$ could be a sequence of random variables satisfying a recurrence relation of the type

$$X_n \stackrel{d}{=} f(X_{n-1}, X_{n-2}, \ldots, X_{n-k}), \quad n \geq k,$$

where the symbol $\stackrel{d}{=}$ means 'equal in distribution' and the distribution of X_n is the same as the distribution of $f(X_{n-1}, X_{n-2}, \ldots, X_{n-k})$. The function $f(\cdot)$ is fixed and known. Typically, the function $f(\cdot)$ is of the form

$$f(X_{n-1}, X_{n-2}, \ldots, X_{n-k}) = a_1 X_{n-1} + a_2 X_{n-2} + \cdots + a_k X_{n-k} + Y_n,$$

where $a_1, a_2, \ldots, a_k$ are constants and Y_n is a sequence of random variables independent of the sequence $X_n, n \geq 0$. The goal is to determine the distribution and properties of X_n. In order to determine the distribution and properties of X_n such as mean and variance, we ought to have a knowledge of the distribution of the initial random variables $X_0, X_1, \ldots, X_{k-1}$, and also that of $Y_n, n \geq k$.

The probabilistic recurrence relation outlined above looks like an Autoregressive model encountered in Time Series Analysis but the random variables involved have special distributions. In our 'modus operandi' the random variables have discrete distributions. Further, we do discuss situations in which the recurrence relations are not exactly the type outlined above but in spirit they are similar.

In this article, we present a number of examples from discrete probability which are unified by the presence of a probabilistic recurrence relation. Some of these problems came from our consulting work. These are listed below:

1. Dermal Patch Problem
2. Patterns in Coin Tossing
3. Chemical Bonding Problem
4. Yell Game
5. Noodles Problem

These examples can be used in a class-room setting to demonstrate how probabilistic recurrence relations arise naturally in some probabilistic investigations and exhort students to find ways to solve such recurrence relations.

15.2 An Illustrative Example

Let $X_1, X_2, \cdots$ be a sequence of independent identically distributed Bernoulli random variables with $Pr(X_1 = 1) = p = 1 - Pr(X_1 = 0)$ with $0 < p < 1$ fixed. What is the distribution of $S_n = X_1 + X_2 + \cdots + X_n$? There are many ways this problem can be solved. We can solve this problem by building a probabilistic recurrence relation. Note that $S_n = S_{n-1} + X_n$. We postulate the induction hypothesis that $S_n \sim \text{Binomial}(n, p)$ for all $n \geq 1$, i.e., $Pr(S_n = r) = \binom{n}{r} p^r (1-p)^{n-r}$, $r = 0, 1, \cdots, n$. The result is true for $n = 1$. Assume that the result is true for all $k \leq n$. We will show that the result is true also for $k = n + 1$. Observe that $S_{n+1} = 0$ if and only if $S_n = 0$ and $X_{n+1} = 0$. In view of independence,

$$Pr(S_{n+1} = 0) = Pr(S_n = 0) \cdot Pr(X_{n+1} = 0) = (1-p)^{n+1}.$$

Let $1 \leq r \leq n$. Observe that $S_{n+1} = r$ if and only if $S_n = r$ and $X_{n+1} = 0$ or $S_n = r - 1$ and $X_{n+1} = 1$. In view of independence,

$$\begin{aligned} Pr(S_{n+1} = r) &= Pr(S_n = r) \cdot Pr(X_{n+1} = 0) + Pr(S_n = r-1) \cdot Pr(X_{n+1} = 1) \\ &= \binom{n}{r} p^r (1-p)^{n-r} \cdot (1-p) + \binom{n}{r-1} p^{r-1} (1-p)^{n-r+1} \cdot p \\ &= \left(\binom{n}{r} + \binom{n}{r-1} \right) p^r (1-p)^{n-r+1} \\ &= \binom{n+1}{r} p^r (1-p)^{n-r+1}. \end{aligned}$$

Finally, $S_{n+1} = n + 1$ if and only if $S_n = n$ and $X_{n+1} = 1$. Consequently,

$$Pr(S_{n+1} = n+1) = Pr(S_n = n) \cdot Pr(X_{n+1} = 1) = p^n \cdot p = p^{n+1}.$$

Thus $S_{n+1} \sim \text{Binomial}(n+1, p)$.

15.3 Dermal Patch Problem

A dermal patch has been created to release doses incrementally over a period of 10 days into an organism to which the patch has been attached. The goal of the doses is to reach ten receptors, currently dormant, inside the organism, and make

them active. On Day 1, the patch releases one dose, which latches onto one receptor and makes it active. On Day 2, the patch releases two doses, which latch onto two receptors, one dose per receptor. If the receptor is already active, the new dose makes it inactive. If the receptor is not active, the new dose makes it active. On Day 3, the patch releases three doses of medicine, which latch onto three receptors, one dose per receptor. The active or inactive status of the receptor which received a dose follows the same paradigm outlined above. This will continue until the tenth day, on which ten doses are released, which latch onto all the ten receptors, one dose per receptor. The question, posed by a pharmaceutical company, is to determine how many receptors are expected to be active at the end of the tenth day.

We have generalized and synthesized the problem. We have christened the generalization as the 'Light Bulb' problem.

In the light bulb problem, we assume that there are $n \geq 1$ light bulbs, serially numbered from 1 to n, each one attached to a toggle switch. If the bulb is off and its switch is pressed, the bulb lights up. If the bulb is on and its switch is pressed, then the bulb is off. Initially all bulbs are set off. On Day $i (= 1, 2, \ldots, n)$ exactly i of the n switches are randomly pressed. The question raised is to determine the expected number of light bulbs on at the end of the nth day. The problem was first considered in [8].

Let X_n be the number of light bulbs on at the end of day n. Write

$$X_n = X_{n1} + X_{n2} + \cdots + X_{nn},$$

where $X_{ni} = 1$, if the ith bulb is on at the end of Day n, and $= 0$ otherwise. Note that $X_{n1}, X_{n2}, \cdots, X_{nn}$ are exchangeable. It suffices to calculate $p_1^{(n)} = Pr(X_{n1} = 1)$. This would give $E(X_n) = np_1^{(n)}$. Let

$$p_1^{(r)} = \text{ Probability that the bulb attached to Switch 1 is on at the end of Day } r.$$

Then the light bulb attached to Switch 1 lights up at the end of Day r if and only if (1) the light bulb was on at the end of Day $(r-1)$ and Switch 1 was not pressed on the rth day; or (2) the light bulb was off at the end of Day $(r-1)$ and Switch 1 was not pressed on the rth day. We then come up with the following recurrence relation, a non-homogeneous difference equation of order 1,

$$p_1^{(r)} = \left(1 - \frac{r}{n}\right) p_1^{(r-1)} + \frac{r}{n}(1 - p_1^{(r-1)}) = \frac{r}{n} + \left(1 - \frac{2r}{n}\right) p_1^{(r-1)}, r = 2, 3, \cdots, n,$$

with boundary condition $p_1^{(1)} = 1/n$. To solve this equation, letting $q_1^{(r)} = p_1^{(r)} - \frac{1}{2}$, $r = 1, 2, \cdots, n$, we have the following recurrence relation

$$q_1^{(r)} = \left(1 - \frac{2r}{n}\right) q_1^{(r-1)},$$

for $r = 2, 3, \cdots, n$ with boundary condition $q_1^{(1)} = \frac{1}{n} - \frac{1}{2}$. Thus, we obtain

$$p_1^{(r)} = \frac{1}{2} + \left(\frac{1}{n} - \frac{1}{2}\right) \prod_{i=2}^{r} \left(1 - \frac{2i}{n}\right) = \frac{1}{2}\left(1 - \prod_{i=1}^{r} \left(1 - \frac{2i}{n}\right)\right),$$

for $r = 1, 2, \cdots, n$. Therefore,

$$E(X_n) = \frac{n}{2}\left(1 - \prod_{i=1}^{n} \left(1 - \frac{2i}{n}\right)\right).$$

Note that $EX_n = \frac{n}{2}$ if n is even, and $\approx \frac{n}{2}$ if n is odd. The idea of recurrence relation has been extensively exploited in [8] to calculate the variance, covariance and the distribution of X_n.

15.4 Patterns in Coin Tossing

If we toss a fair coin four times, what are the chances of getting the pattern HHHH? HHHT? They are the same. Let us pose the problem differently. Keep tossing the coin until we get the pattern HHHH. How many tosses are needed to get the pattern HHHH? How many tosses are needed to get the pattern HHHT? It is harder to get the pattern HHHH than the pattern HHHT in the sense that the expected number of tosses required to get the pattern HHHH is greater than the corresponding expectation for HHHT.

Questions on expected number of tosses needed for any pattern of outcomes to appear for the first time have been solved elegantly using martingale arguments in [6] and [5]. Their work was further extended in [7]. In this section, we focus on the distribution of the number of tosses required. We show how probabilistic recurrence relations help solve the problem and provide a rich source and variety of number sequences.

Let T_{H} be the number of tosses required to get heads for the first time. It is an elementary exercise that the distribution of T_{H} is geometric, i.e.,

T_{H}:	1	2	3	$\cdots$
Probability:	$1/2$	$(1/2)^2$	$(1/2)^3$	$\cdots$

with $E(T_{\mathrm{H}}) = 2$.

Let T_{HH} be the number of tosses required to get the pattern HH. It is not hard to show that its distribution is given by

T_{HH}:	2	3	4	5	6	$\cdots$
Probability:	$(1/2)^2$	$(1/2)^3$	$2(1/2)^4$	$3(1/2)^5$	$5(1/2)^6$	$\cdots$

with $E(T_{HH}) = 6$. A surprising feature about the distribution is that the coefficients of power of $1/2$ form the Fibonacci sequence! The distribution of T_{HH} can be derived from that of T_H. The key idea is a probabilistic recurrence relation.

Suppose that we are observing T_H. Look at the $(T_H + 1)^{st}$ toss. It could be either H with probability 1/2 or T with probability 1/2. If it is H, we have the pattern HH occurring for the first time and $T_{HH} = 1 + T_H$. If it is T, we need to start from scratch beyond the $(T_H + 1)^{st}$ toss to look for the pattern HH. Consequently, $T_{HH} = 1 + T_H + T^*_{HH}$, where T^*_{HH} is an independent copy of T_{HH}. These deliberations lead to the following probabilistic recurrence relation

$$T_{HH} = \begin{cases} 1 + T_H & \text{with probability 1/2,} \\ 1 + T_H + T^*_{HH} & \text{with probability 1/2.} \end{cases}$$

It is easy to exploit this recurrence relation to get ET_{HH}.

$$\begin{aligned} ET_{HH} &= (1 + ET_H) \cdot \left(\frac{1}{2}\right) + (1 + ET_{HH} + ET^*_{HH}) \cdot \left(\frac{1}{2}\right) \\ &= 3 + \left(\frac{1}{2}\right) ET_{HH}, \end{aligned}$$

implying that $ET_{HH} = 6$. This relation can be exploited to calculate variance of T_{HH} and distribution of T_{HH}. As for the distribution of T_{HH}, note that

$$T_{HH} = 2 \quad \text{if and only if} \quad T_H = 1.$$

Consequently,

$$P(T_{HH} = 2) = \left(\frac{1}{2}\right) Pr\,(T_H = 1) = \left(\frac{1}{2}\right)^2.$$

Also,

$$T_{HH} = 3 \quad \text{if and only if} \quad T_H = 2.$$

Hence,

$$P(T_{HH} = 3) = \left(\frac{1}{2}\right) Pr\,(T_H = 2) = \left(\frac{1}{2}\right)^3.$$

Now

$$\begin{aligned} T_{HH} = 4 \quad &\text{if and only if} \quad T_H = 3 \quad \text{or} \\ &T_{HH} = 1 \quad \text{and} \quad T^*_{HH} = 2. \end{aligned}$$

Therefore,

$$\begin{aligned} Pr\,(T_{HH} = 4) &= Pr\,(T_{HH} = 3) \left(\frac{1}{2}\right) + Pr\,(T_{HH} = 1) Pr\,(T^*_{HH} = 2) \left(\frac{1}{2}\right) \\ &= \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right) = 2\left(\frac{1}{2}\right)^4. \end{aligned}$$

One can proceed inductively to lay out the entire distribution of T_{HH}. It should be pointed out that there are different ways to develop a recurrence relation for T_{HH}. Following the same trail, we have obtained the distribution of T_{HHH}.

T_{HHH}:	3	4	5	6	7	8	9	$\cdots$
Probability:	$(1/2)^3$	$(1/2)^4$	$2(1/2)^5$	$4(1/2)^6$	$7(1/2)^7$	$13(1/2)^8$	$24(1/2)^9$	$\cdots$

A serendipitous discovery is that the coefficients of powers of $\left(\frac{1}{2}\right)$ in the distribution constitute a tribonacci sequence, i.e.,

$$f_1 = 1, f_2 = 1, f_3 = 2, f_4 = 4, f_5 = 7, f_6 = 13,$$
$$f_7 = 24, \cdots, f_n = f_{n-1} + f_{n-2} + f_{n-3},$$

etc. In general, one can show that the coefficients of powers of $\left(\frac{1}{2}\right)$ in the distribution of $T_{\underbrace{HH\cdots H}_{n}}$ constitute a n-bonacci sequence. Further,

$$E(T_{\underbrace{HH\cdots H}_{n}}) = 2^{n+1} - 2.$$

Using similar ideas, we have obtained the distribution of T_{HT}.

T_{HT}:	2	3	4	5	$\cdots$
Probability:	$(1/2)^2$	$2(1/2)^3$	$3(1/2)^4$	$4(1/2)^5$	$\cdots$

with $E(T_{\text{HT}}) = 4$. Note that the coefficients of powers of $\left(\frac{1}{2}\right)$ in the distribution constitute the entire sequence of natural numbers.

A very rich family of distributions on $\{0, 1, 2, 3, \cdots, \}$ can be generated by looking at the distributions of T_{pattern} for a variety of patterns. Further, by looking at the coefficients of powers of $\left(\frac{1}{2}\right)$ in the distributions, we will have a rich variety of positive integer sequences. In [4], expectations for patterns of length 3 are given using Markov Chain ideas. Here is a partial list of expectations.

Pattern	Expectation
H	2
HH	6
HT	4
HHH	14
HHT	8
HTH	10
THH	8

One can provide a general formula for $E(T_{\text{pattern}})$ for any given 'Pattern' of heads and tails. If the 'Pattern' is of length n, then

$$E(T_{\text{pattern}}) = 2^n + b_1 \cdot 2^{n-1} + b_2 \cdot 2^{n-2} + \cdots + b_{n-1} \cdot 2,$$

for some $b_1, b_2, \cdots, b_{n-1}$ in $\{0, 1\}$. The following algorithm gives the coefficients. Let $a_1 a_2 \cdots a_n$ be a pattern of length n.

Stage 0: Write the pattern as follows.

$$\begin{array}{l} a_1 a_2 \cdots a_n \\ a_1 a_2 \cdots a_n \end{array}$$

There is a perfect alignment.

Stage 1: Slide the second pattern one position to the right.

$$\begin{array}{l} a_1 a_2 a_3 \cdots a_n \\ \quad a_1 a_2 \cdots a_{n-1} a_n \end{array}$$

If there is a perfect match in the common positions of both the patterns, $b_1 = 1$, otherwise, $b_1 = 0$.

Stage 2: Slide the second pattern two positions to the right.

$$\begin{array}{l} a_1 a_2 a_3 a_4 \cdots a_n \\ \quad\quad a_1 a_2 \cdots a_{n-2} a_{n-1} a_n \end{array}$$

If there is a perfect match in the common positions of both the patterns, $b_2 = 1$, otherwise, $b_2 = 0$, and so on ....

As an illustration, look at the pattern HHTTHH.

Stage 0: Write the pattern as follows.

$$\begin{array}{l} HHTTHH \\ HHTTHH \end{array}$$

There is a perfect alignment.

Stage 1: Slide the second pattern one position to the right.

$$\begin{array}{l} H\ \ HTTHH \\ \quad HHTTHH \end{array}$$

$b_1 = 0$.

Stage 2: Slide the pattern two positions to the right.

$$\begin{array}{l} HH\ \ TTHH \\ \quad\quad HHTTHH \end{array}$$

$b_2 = 0$

Stage 3: Slide the pattern three positions to the right.

$$\begin{array}{l} HHT\ \ THH \\ \qquad\ HHTTHH \end{array}$$

$b_3 = 0$

Stage 4: Slide the pattern four positions to the right.

$$\begin{array}{l} HHTT\ \ HH \\ \qquad\ HHTTHH \end{array}$$

$b_4 = 1$

Stage 5: Slide the pattern five positions to the right.

$$\begin{array}{l} HHTTH\ \ H \\ \qquad\ HHTTHH \end{array}$$

$b_5 = 1$

Consequently,

$$E(T_{\text{HHTTHH}}) = 2^6 + 2^2 + 2^1 = 70.$$

15.5 Chemical Bonding Problem

A certain system has ten molecules in some hierarchical order from 1 to 10, say. A catalyst is added to the system and the resultant reaction lines up the molecules. One can read the hierarchies of the molecules from left to right. If there are molecules in the line-up in consecutive increasing hierarchical order, then these molecules merge to become one molecule. A new hierarchical order sets in among the molecules that are left over including the merged ones. Again, a catalyst is added to the system and the whole process of merging renews again. The process continues until we get a single molecule. The question of interest is how many times catalysts are expected to be added until we get a single molecule.

We have generalized and synthesized the problem. We relabel the problem as 'Card Shuffling Problem'. (A related problem has been presented in [2]).

In the card shuffling problem, we assume that n cards, serially numbered from 1 to n, are shuffled thoroughly and spread out in a line on a table facing up. The cards were then read from left to right, and those cards with consecutive numbers in increasing order are merged. After the mergers, the cards are renumbered serially from 1 onwards to whatever number of cards available now. Cards are shuffled thoroughly again and laid out on a table in a line facing up, etc. The process continues until only one card is left. See [9] for more details.

Let X_n be the number of shuffles needed so that there is only one card left at the end. The principal questions are: $E(X_n)$? Var(X_n)? Distribution of X_n? We will start with a special case $n = 2$ to illustrate our approach to the problem, and then extend the idea to the general case.

Assume $n = 2$, let $\tilde{X}_2$ be an independent copy of X_2. X_2 takes values $\{1, 2, 3, \cdots\}$. If the first shuffle ends up with the permutation $\underline{1\ 2}$ (with probability $1/2$), only one shuffle is needed. However, if the first shuffle ends with the permutation $\underline{2\ 1}$ (with probability $1/2$), then the cards are needed to be shuffled again. The number of shuffles needed to have one card left is $\tilde{X}_2$. In summary, we have the following.

Possible permutations:	Probability	Value of X_2
$\{1\ 2\}$	$1/2$	1
$\{2\ 1\}$	$1/2$	$1 + \tilde{X}_2$

In other words,

$$X_2 = \begin{cases} 1 & \text{with prob. 1/2, corresponding to the shuffle result } \{1\ 2\}, \\ 1 + \tilde{X}_2 & \text{with prob. 1/2, corresponding to the shuffle result } \{2\ 1\}, \end{cases}$$

This is a probabilistic recurrence relation. It can be exploited to calculate $E(X_2)$. For example

$$E(X_2) = 1 * (1/2) + E(1 + \tilde{X}_2) * (1/2) = 1 + \frac{1}{2}E(X_2),$$

implying that $E(X_2) = 2$.

Before moving to the general case, we follow the notation from [2, p. 168], by letting $T_{n,s}$ be the number of permutations of $\{1, 2, \cdots, n\}$ such that in each permutation there are s pairs of consecutive increasing integers in the permutation. For $n \leq 8$, $T_{n,s}$ is given by Table 1 on the same page. We further let $A(n, k), k = 1, 2, \cdots, n$ be the number of permutations of $\{1, 2, \cdots, n\}$ in which k cards are left after merging from n cards. It is shown in [9] that $A(n, k) = T(n, n - k)$ and $\sum_{k=1}^{n} A(n, k) = n!$.

Suppose that there are n cards initially. Let $\tilde{X}_k$ be an independent copy of X_k, $k = 2, 3, \cdots, n$. If the first shuffle ends up with the permutation $\underline{1\ 2 \cdots n}$ (with probability $A(n, 1)/n! = 1/n!$), only one shuffle is needed. In general, if there are k cards left (with probability $A(n, k)/n!, k = 2, 3, \cdots$) after the first shuffle and all possible mergers, then the k cards are needed to be shuffled again and the number of shuffles needed to have one card left is $\tilde{X}_k$. Therefore, we have the following recurrence relation.

$$X_n = \begin{cases} 1 & \text{with probability } A(n, 1)/n!, \\ 1 + \tilde{X}_2 & \text{with probability } A(n, 2)/n! \\ \cdots & \cdots \\ 1 + \tilde{X}_n & \text{with probability } A(n, n)/n! \end{cases}$$

Hence, we have

$$E(X_n) = 1 + \frac{A(n,2)}{n!}E(X_2) + \cdots + \frac{A(n,n-1)}{n!}E(X_{n-1}) + \frac{A(n,n)}{n!}E(X_n),$$

and so

$$E(X_n) = \left(1 + \sum_{k=2}^{n-1} \frac{A(n,k)}{n!}E(X_k)\right) \Big/ \left(1 - \frac{A(n,n)}{n!}\right).$$

The following table gives $E(X_n)$ when $n \leq 10$. More results can be found in [9].

n	2	3	4	5	6	7	8	9	10
$E(X_n)$	2	3.33	4.61	5.84	7.02	8.17	9.30	10.42	11.40

15.6 Yell Game

This game is played by many improved comedy troupes for encouraging cooperation between performers. See [1] for details. The game is played by N individuals who stand in a circle, initially looking down at the floor. On a signal, all players simultaneously lift their gaze from the floor and each player fixes his/her gaze exactly at one other player in the circle. If any two players make eye contact with each other, they yell at each other and they are out of the game. The game continues with players who remain. In [1], the focus was on calculating probability of silence. In this section, we focus on determining the distribution of the number of people left over at each turn of the game.

Let $n \geq 2$. Let $X_0^{(n)} = n =$ the number of players at the start of the game. Let $X_k^{(n)} =$ the number of players at the end of kth turn of the game. If n is even, the possible values of $X_k^{(n)}$ are $0, 2, \cdots, n$. If n is odd, the possible values of $X_k^{(n)}$ are $1, 3, \cdots, n$. Let us look at some special cases. Suppose $n = 2$. Then $X_0^{(2)} = 2, X_1^{(2)} = 0, X_2^{(2)} = 0, \cdots$ etc. Suppose $n = 3$. The distribution of $X_1^{(3)}$ is given by

$X_1^{(3)}$:	1	3
Probability	3/4	1/4

It is clear that $X_k^{(3)}, k \geq 0$ constitutes a Markov Chain with state space {1, 3} and transition probability matrix

$$P = \begin{pmatrix} 1 & 0 \\ 3/4 & 1/4 \end{pmatrix}.$$

Using Markov Chain ideas or a recurrence relation approach, one can show that the distribution of $X_k^{(3)}$ is

$X_k^{(3)}$:	1	3
Probability	$\frac{4^k-1}{4^k}$	$\frac{1}{4^k}$

with $E(X_k^{(3)}) = 1 + \frac{2}{4^k}$. Let us look at the case $n = 4$. The distribution of $X_1^{(4)}$ is given by

$X_1^{(4)}$:	0	2	4
Probability	1/27	8/27	18/27

and $E(X_1^{(4)}) = 88/27$. The sequence $X_k^{(4)}, k \geq 0$ constitutes a Markov Chain with state space $\{0, 2, 4\}$ and transition probability matrix

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1/27 & 8/27 & 18/27 \end{pmatrix}.$$

Using Markov Chain ideas or a recurrence relation approach, one can show that the distribution of $X_k^{(4)}$ is

$X_k^{(4)}$:	0	2	4
Probability	$1 - \frac{2^k \cdot 13}{3^{k+2}}$	$\frac{2^k \cdot 4}{3^{k+2}}$	$\frac{2^k \cdot 9}{3^{k+2}}$

with $E(X_k^{(4)}) = \frac{2^{k+2} \cdot 11}{3^{k+2}}$. We do not have a general formula for the distribution of $X_k^{(n)}$ for any n. One interesting problem is to find an expression for the probability of 'silence' for any n. More precisely, let

$$\begin{aligned} p_n = Pr\,(X_1^{(n)} = n) &= Pr\,(\text{all players stay in the game at the end of the first turn}) \\ &= Pr\,(\text{No yells}). \end{aligned}$$

It will be interesting to obtain an asymptotic formula for p_n.

15.7 Noodles Problem

15.7.1 Introduction

There are n strings. Each string has two ends. Out of $2n$ ends two ends are selected at random and tied. If the two ends come from the same string, we will have a loop

right away. Otherwise, we will have $(n-1)$ strings with $2(n-1)$ ends. We now select two ends at random from these $2(n-1)$ strings. The selection will continue until the last two ends available are tied. Let X_n be the total number of loops that will form. The possible values of X_n are $1, 2, \cdots, n$. The objective of this section is to determine the distribution of X_n, its expected value, the variance, and the asymptotic normality.

15.7.2 Expected Value

The random variables X_{n-1} and X_n have a connection. If the initial selection of two ends come from the same string, we will have a loop right away and $X_n = 1 + X_{n-1}$. On the other hand, if the initial selection of two ends come from two different strings, then $X_n = X_{n-1}$. The probability that two ends come from the same string is given by $1/(2n-1)$. The complementary event has the probability of $2(n-1)/(2n-1)$. Formally, the relationship between the random variables can be written as

$$X_n \stackrel{d}{=} \begin{cases} 1 + X_{n-1} & \text{with prob. } 1/(2n-1), \\ X_{n-1} & \text{with prob. } 2(n-1)/(2n-1). \end{cases} \tag{15.1}$$

This fundamental recurrence relation in distribution can provide formulas for $E(X_n)$ and $Var(X_n)$. In addition, the distribution of X_n is derivable. First, let us work with the expected value. Observe that

$$\begin{aligned} E(X_n) &= \frac{1}{2n-1} E(1 + X_{n-1}) + \frac{2(n-1)}{2n-1} E(X_{n-1}) \\ &= \frac{1}{2n-1} + E(X_{n-1}), n = 2, 3, \cdots. \end{aligned}$$

The boundary condition is $E(X_1) = 1$. The formula for the expected value can be set out directly from the recurrence relation between two consecutive expectations.

$$E(X_n) = 1 + \frac{1}{3} + \frac{1}{5} + \cdots + \frac{1}{2n-1}.$$

Note that the expected value diverges to ∞, rather slowly, at the pace of $\ln n$. An approximation for the expected value is given by

$$E(X_n) = \frac{\ln n}{2} + \ln 2 + \frac{1}{2}\gamma,$$

where $\gamma \sim 0.5772$ is the Euler's constant. This approximation is good with agreement between the exact value and the approximation taking place in the first two decimal places for n equal to as low as 15.

15.7.3 Variance

The recurrence relation between the random variables X_n and X_{n-1} can be exploited to get a recurrence relation between $Var(X_n)$ and $Var(X_{n-1})$. However, we present another relation between X_n and X_{n-1} by introducing a Bernoulli random variable. In the context of n strings, let

$$I_n = \begin{cases} 1 & \text{if the randomly chosen ends come from the same string,} \\ 0 & \text{otherwise.} \end{cases} \tag{15.2}$$

The probability distribution of I_n is given by

$$P(I_n = 1) = \frac{1}{2n-1} = 1 - P(I_n = 0).$$

The distributional recurrence relation (15.1) is equivalent to

$$X_n \stackrel{d}{=} X_{n-1} + I_n \tag{15.3}$$

with X_{n-1} and I_n independently distributed. Consequently, in view of independence of X_{n-1} and I_n,

$$Var(X_n) = Var(X_{n-1}) + Var(I_n) = Var(X_{n-1}) + \frac{2(n-1)}{(2n-1)^2}, n = 2, 3, \cdots.$$

The boundary condition is given by $Var(X_1) = 0$. The formula for the variance works out to be

$$Var(X_n) = 1 + \frac{1}{3} + \frac{1}{5} + \cdots + \frac{1}{2n-1} - \left(1 + \frac{1}{3^2} + \frac{1}{5^2} + \cdots + \frac{1}{(2n-1)^2}\right).$$

The variance also diverges to ∞, rather slowly, at the pace of $\ln n$. An approximation to the variance is given by

$$Var(X_n) = \frac{\ln n}{2} + \ln 2 + \frac{1}{2}\gamma - \frac{3}{4}\frac{\pi^2}{6}.$$

15.7.4 Distribution

The recurrence relation (15.1) and (15.3) can be used for obtaining the distribution of X_n from that of X_{n-1}. Let the distribution of X_{n-1} be given by

X_{n-1}	1	2	$\cdots$	$n-1$
Probability	$p_{n-1,1}$	$p_{n-1,2}$	$\cdots$	$p_{n-1,n-1}$.

Let us denote the probability distribution of X_{n-1} by the $(n-1)$-tuple: $(p_{n-1,1}, p_{n-1,2}, \cdots, p_{n-1,n-1})$. From (15.2), following the same notation, the probability distribution of X_n is given by

$$(p_{n,1}, p_{n,2}, \cdots, p_{n,n}) = \frac{1}{2n-1} * (0, p_{n-1,1}, p_{n-1,2}, \cdots, p_{n-1,n-1}) + \frac{2n-2}{2n-1} * (p_{n-1,1}, p_{n-1,2}, \cdots, p_{n-1,n-1}, 0).$$

The distribution of X_n is a convex combination of two probability n-tuples created from the distribution of X_{n-1}. From this, it follows that, for $n = 2, 3, \cdots$,

$$p_{n,1} = \frac{2n-2}{2n-1} p_{n-1,1} = \frac{(2n-2)(2n-4)\cdots(4)(2)}{(2n-1)(2n-3)\cdots(5)(3)};$$
$$p_{n,j} = \frac{1}{2n-1} p_{n-1,j-1} + \frac{2n-2}{2n-1} p_{n-1,j},\ j = 2, 3, \cdots, n-1;$$
$$p_{n,n} = \frac{1}{2n-1} p_{n-1,n-1} = \frac{1}{(2n-1)(2n-3)\cdots(5)(3)}.$$

The distribution of X_n for selected values of n is tabulated below.

	The distribution of X_n					
n	1	2	3	4	5	6
1	1	0	0	0	0	0
2	$\frac{2}{3}$	$\frac{1}{3}$	0	0	0	0
3	$\frac{8}{15}$	$\frac{6}{15}$	$\frac{1}{15}$	0	0	0
4	$\frac{48}{105}$	$\frac{44}{105}$	$\frac{12}{105}$	$\frac{1}{105}$	0	0
5	$\frac{384}{945}$	$\frac{400}{945}$	$\frac{140}{945}$	$\frac{20}{945}$	$\frac{1}{945}$	0
6	$\frac{3840}{10395}$	$\frac{4384}{10395}$	$\frac{1800}{10395}$	$\frac{340}{10395}$	$\frac{30}{10395}$	$\frac{1}{10395}$

15.7.5 Asymptotic Normality

With the recurrence relation (15.3), we can derive the asymptotic normality for X_n. Note that, X_n can be expressed in terms of the sum of n independent but not identically distributed Bernoulli random variables, i.e.,

$$X_n = I_1 + I_2 + \cdots + I_n,$$

where I_m is a Bernoulli random variable with the distribution given by

$$P(I_m = 1) = \frac{1}{2m-1} = 1 - P(I_m = 0), m = 1, 2, \cdots, n.$$

Therefore, we have the following asymptotic normality.

Theorem 15.1. *As* $n \to \infty$, $\frac{X_n - E(X_n)}{\sqrt{Var(X_n)}} \xrightarrow{d} N(0,1)$.

Proof. We consider the normalized random variable $Z_{n,m} = \frac{I_m - E(I_m)}{\sqrt{Var(X_n)}}$. Then $E(Z_{n,m}) = 0$ and $\sum_{m=1}^{n} EZ_{n,m}^2 = 1$. In addition, since $VarX_n = O(\ln n)$, for any fixed $\varepsilon > 0$, taking sufficiently large n, we have $(\ln n)^{-1/2} < \varepsilon$, which implies that, for any fixed $\varepsilon > 0$,

$$\sum_{m=1}^{n} E(|Z_{n,m}|^2; |Z_{n,m}| > \varepsilon) \to 0,$$

as $n \to \infty$. By the Lindeberg–Feller Central Limit theorem, for example, p. 116 in [3], the asymptotic normality holds.

15.8 Conclusions

There is a plethora of discrete probability problems coming from a variety of research disciplines. We have presented a sample of these problems. We have developed a probabilistic recurrence relation methodology to solve these problems. There are some problems remain unsolved.

Acknowledgements H. Zhang would like to thank the Department of Environmental Health at the University of Cincinnati for the support and splendid hospitality during his visits.

References

1. Balas B, Connor C (2004–2005) Look up and scream: Analytical difficulties in improv comedy. J Recreat Math 33:32–38
2. David F, Barton D (1962) Combinatorial chance. Hafner Publishing Company, New York
3. Durrett R (1996) Probability: Theory and examples, 2nd ed. Duxbury, Belmont
4. Durrett R (1999) Essentials of stochastic processes. Springer, New York
5. Gerber H, Li S (1981) The occurrence of sequence patterns in repeated experiments and hitting times in Markov chains. Stoch Process Appl 11:101–108
6. Li S (1980) A martingale approach to the study of occurence of sequence patterns in repeated experiments. Ann Prob 8:1171–1176

7. Pozdnyakov V, Kulldorff M (2006) Waiting times for patterns and a method of gambling teams. Am Math Mon 113:134–143
8. Rao CR, Rao MB, Zhang H (2007) One bulb? Two bulbs? How many bulbs light up? A discrete probability problem involving dermal patches. Sankhyā 69:137–161
9. Rao MB, Zhang H, Huang C, Chen F (2009) A discrete probability problem in bonding of molecules. Preprint

Chapter 16
On Some Inequalities of Chernoff–Borovkov–Utev Type for Circular Distributions

B.L.S. Prakasa Rao

Abstract We discuss some classical inequalities such as Wirtinger inequality and weighted Wirtinger type inequality for 2π-periodic functions and study their applications for obtaining Chernoff–Borovkov–Utev type inequalities for probability distributions with support $[0, 2\pi]$. In addition we derive Chernoff type inequalities for the wrapped normal distribution and von-Mises distribution.

16.1 Introduction

Let X be a standard normal random variable and g$(\cdot)$ be an absolutely continuous function. The inequality

$$\mathrm{Var}[g(X)] \leq E([g'(X)]^2) \tag{16.1}$$

was first proved by [9] and later rediscovered by [4] as a special case of a general result for log-concave densities and was obtained independently by [5]. Analogous inequalities for several discrete and continuous distributions are known. A matrix variance inequality for the normal distribution was recently obtained by [10]. Matrix variance inequalities for some multivariate distributions are derived in [13]. For a recent overview on some stochastic inequalities and stochastic identities, see [14]. Borovkov and Utev [3] obtained a general inequality of the type (16.1). We will discuss similar inequalities for random variables defined on the interval $[0, 2\pi]$. It is known that the concept of mean and variance for circular random variables do not have the same interpretation as they do for random variables taking values in the real line. Statistical methods for analysis of directional or angular data are of great interest. Following [6], we derive some inequalities for circular variance for circular random variables and obtain some inequalities for standard wrapped normal

B.L.S. Prakasa Rao
Department of Mathematics and Statistics, University of Hyderabad, Hyderabad 500 046, India
e-mail: blsprao@gmail.com

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_16,

distribution and standard von-Mises distribution. Derivation of bounds on moments, in particular on variances, for functions of circular random variables is of interest in directional data analysis as the exact distributions of these random variables might be difficult to compute as in the usual case of functions of random variables taking values in the real line. The variance bounds for functions of circular random variables derived in Sect. 16.3 deal with the general case. However it is possible to derive better bounds for the wrapped normal distribution and von-Mises distribution. These are discussed in Sects. 16.4 and 16.5 respectively. We now discuss some classical inequalities from analysis such as Hardy's inequalities with weights, Wirtinger inequality and weighted Wirtinger type inequality which are also related to the problems we are investigating.

16.1.1 Hardy's Inequalities with Weights

Hardy's inequality (cf. [16, p. 20]) states that if p and b satisfy the conditions $1 \leq p < \infty$ and $bp < -1$, then

$$\left[\int_0^\infty \left|x^b \int_0^x f(t)dt\right|^p dx\right]^{1/p} \leq \frac{-p}{bp+1}\left[\int_0^\infty \left|x^{b+1} f(x)\right|^p dx\right]^{1/p} \tag{16.2}$$

and the constant $\frac{-p}{bp+1}$ is the best possible. Muckenhoupt [8] generalized this inequality and proved the following result.

Theorem 16.1. *Suppose μ and ν are Borel measures and $1 \leq p < \infty$. Then there is a finite constant C for which*

$$\left[\int_0^\infty \left|\int_0^x f(t)dt\right|^p d\mu(x)\right]^{1/p} \leq C\left[\int_0^\infty |f(x)|^p d\nu(x)\right]^{1/p} \tag{16.3}$$

if and only if

$$B = \sup_{r>0}\left\{[\mu([r,\infty))]^{1/p}\left[\int_0^r \left(\frac{dv}{dx}\right)^{-q/p} dx\right]^{1/q}\right\} \tag{16.4}$$

where $\frac{dv}{dx}$ denotes the Radon–Nikodym derivative of the absolutely continuous component of ν with respect to the Lebesgue measure and q is such that $\frac{1}{p} + \frac{1}{q} = 1$. Furthermore, if C is the least constant for which (16.3) is true, then

$$B \leq C \leq p^{1/p} 1^{1/q} B \tag{16.5}$$

for $p > 1$ and $B = C$ for $p = 1$.

An extensive discussion on inequalities of this type and related inequalities known as the logarithmic Sobolev inequalities is given in [2].

The following discussion deals with inequalities for 2π-periodic absolutely continuous functions.

16.1.2 Wirtinger Inequality

Suppose g is a 2π-periodic and absolutely continuous function with $g' \in L^2[0, 2\pi)$ where g' denotes the derivative of g whenever it exists. Further suppose that

$$\int_0^{2\pi} g(\theta)d\theta = 0.$$

Then it is known that

$$\int_0^{2\pi} [g(\theta)]^2 d\theta \leq \int_0^{2\pi} [g'(\theta)]^2 d\theta$$

equality occurring if and only if $g(\theta) = A\cos(\theta - \theta_0)$ for some constant A and $0 \leq \theta_0 \leq 2\pi$. This inequality is known as *Wirtinger inequality*. This can be derived easily in the following manner.

Suppose $\{a_n\}$ denotes the sequence of complex Fourier coefficients of the function g. Then those of g' are $\{ina_n\}$. Applying Parseval's identity to g' and observing that

$$a_0 = \frac{1}{2\pi}\int_0^{2\pi} g(\theta)d\theta = 0,$$

we get that

$$\frac{1}{2\pi}\int_0^{2\pi} [g'(\theta)]^2 d\theta = \sum_{n=-\infty}^{\infty} n^2|a_n|^2 \geq \sum_{n=-\infty}^{\infty} |a_n|^2 = \frac{1}{2\pi}\int_0^{2\pi} [g(\theta)]^2 d\theta.$$

This result has been recently extended by [15].

16.1.3 Weighted Wirtinger Type Inequality

Let $C(a, b) > 0$ denote the best constant in the weighted Wirtinger type inequality

$$\int_0^{2\pi} [g(\theta)]^2 a(\theta)d\theta \leq C(a, b)\int_0^{2\pi} [g'(\theta)]^2 b(\theta)d\theta \tag{16.6}$$

where g is a 2π-periodic function and

$$\int_0^{2\pi} g(\theta)a(\theta)d\theta = 0.$$

Let $\mathcal{B}(L)$ be the class of functions $a(\cdot) \in L^{\infty}([0, 2\pi])$ such that $a(\cdot)$ is 2π-periodic and

$$\inf_{0\le\theta\le 2\pi} a(\theta) = 1 \text{ and } \sup_{0\le\theta\le 2\pi} a(\theta) = L.$$

Suppose $a(\theta) = [\gamma(\theta)]^p$ and $b(\theta) = [\gamma(\theta)]^q$ for some $\gamma(\theta) \in \mathcal{B}(M), M > 1$ and $p, q \in R$ such that $p + q \ge 0$. Then the weighted Wirtinger type inequality (16.2) holds with

$$C(a, b) \le \left[\frac{\frac{1}{2\pi}\int_0^{2\pi}[\gamma(\theta)]^{(p-q)/2}d\theta}{\frac{4}{\pi}\arctan(M^{-(p+q)/4})}\right]^2$$

from [15]. Suppose $p = 0 = q$ and $a(\theta) = b(\theta) \equiv 1$. Then it follows that $C(1, 1) \le 1$ leading to the classical Wirtinger inequality. Suppose that $b(\theta) = a(\theta)$ and $p = q = 1$. Then it follows that

$$C(a, a) \le \frac{\pi^2}{16}[\arctan(L^{-1/2})]^2$$

proved in [11].

16.2 Borovkov–Utev Inequality

Suppose ψ is a random variable with a distribution function

$$V(\theta) = \alpha V_1(\theta) + (1 - \alpha)V_2(\theta), 0 \le \theta \le 2\pi$$

where $0 \le \alpha \le 1$ and further suppose that V_1 has the density function $v_1(\theta)$. In addition, suppose that, for some θ_0 in $[0, 2\pi]$ and $c > 0$,

$$\int_{\phi}^{2\pi} (\theta - \theta_0)V(d\theta) \le cv_1(\phi),\ \theta_0 \le \phi \le 2\pi \tag{16.7}$$

and

$$\int_0^{\phi} (\theta_0 - \theta)V(d\theta) \le cv_1(\phi),\ 0 \le \phi \le \theta_0. \tag{16.8}$$

Theorem 16.2. *Suppose the conditions (16.7) and (16.8) hold. Then, for any absolutely continuous function g on $[0, 2\pi]$,*

$$E[g(\psi) - g(\theta_0)]^2 \le \frac{c}{\alpha}E([g'(\psi)]^2). \tag{16.9}$$

The inequality (16.9) follows as a special case of Theorem 1 in [3]. We give the proof here for completeness.

Proof. Observe that

$$
\begin{aligned}
E([g(\psi)-g(\theta_0)]^2) &= \int_0^{2\pi}\left[\int_{\theta_0}^{\theta} g'(\phi)d\phi\right]^2 V(d\theta) \qquad (16.10)\\
&= \int_{\theta_0}^{2\pi}\left[\int_{\theta_0}^{0} g'(\phi)d\phi\right]^2 V(d\theta)\\
&\quad + \int_0^{\theta_0}\left[\int_{\theta}^{\theta_0} g'(\phi)d\phi\right]^2 V(d\theta)\\
&\le \int_{\theta_0}^{2\pi}\left[(\theta-\theta_0)\int_{\theta_0}^{\theta}(g'(\phi))^2 d\phi\right] V(d\theta)\\
&\quad + \int_0^{\theta_0}\left[(\theta_0-\theta)\int_{\theta}^{\theta_0}(g'(\phi))^2 d\phi\right] V(d\theta)\\
&= \int_{\theta_0}^{2\pi}\left[(g'(\phi))^2\int_{\phi}^{2\pi}(\theta-\theta_0)V(d\theta)\right] d\phi\\
&\quad + \int_0^{\theta_0}\left[(g'(\phi))^2\int_0^{\phi}(\theta_0-\theta)V(d\theta)\right] d\phi\\
&\le c\int_{\theta_0}^{2\pi}(g'(\phi))^2 v_1(\phi)d\phi + c\int_0^{\theta_0}(g'(\phi)^2 v_1(\phi)d\phi\\
&= c\int_0^{2\pi}(g'(\phi))^2 v_1(\phi)d\phi\\
&\le \frac{c}{\alpha}E([g'(\psi)]^2).
\end{aligned}
$$

The above inequality also holds for distribution functions on R with support not necessarily compact under conditions similar to those in (16.7) and (16.8) (cf. [3]). As a particular case of Theorem 16.1, the following result can be obtained. □

Theorem 16.3. *Suppose ψ is a random variable with density function $v(\cdot)$ on $[0, 2\pi]$ and there exists θ_0 in $[0, 2\pi]$ and $c > 0$ such that*

$$\int_{\phi}^{2\pi}(\theta-\theta_0)v(\theta)d\theta \le cv(\phi), \theta_0 \le \phi \le 2\pi \qquad (16.11)$$

and

$$\int_0^{\theta}(\theta_0-\theta)v(\theta)d\theta \le cv(\phi), 0 \le \phi \le \theta_0. \qquad (16.12)$$

Then, for any absolutely continuous function $g(\cdot)$ on $[0, 2\pi]$,

$$E[g(\psi)-g(\theta_0)]^2 \le c\, E([g'(\psi)]^2). \qquad (16.13)$$

The following result follows from inequalities derived in [3] following the literature on differential equations.

Theorem 16.4. *Suppose ψ is a random variable with the density function $v(\cdot)$ on $[0, 2\pi]$ and there exists $c > 0$ such that either*

$$g(0) = 0, \int_\theta^{2\pi} v(\phi)d\phi \leq cv(\theta), 0 \leq \theta \leq 2\pi \tag{16.14}$$

or

$$g(2\pi) = 0, \int_0^{\theta} v(\phi)d\phi \leq cv(\theta), 0 \leq \theta \leq 2\pi \tag{16.15}$$

where $g(\cdot)$ is an absolutely continuous function on $[0, 2\pi]$. Then

$$E([g(\psi)]^2) \leq 4c^2 E([g'(\psi)]^2). \tag{16.16}$$

The following two theorems give bounds on the second moments of absolutely continuous functions of random variables defined on the real line.

Theorem 16.5. *Suppose X is a random variable with the distribution function F and the density function $f > 0$ on R and g is an absolutely continuous function with $E([g'(X)]^2) \leq \infty$. Further suppose that*

$$\lim_{x\to\infty} g^2(x)F(x) = \lim_{x\to-\infty} g^2(x)F(x) = 0.$$

Then there exists a positive constant C_g depending on the function g and the distribution function F such that

$$(E[g^2(X)])^2 \leq C_g E([g'(X)]^2).$$

Proof. Applying integration by parts, we get that

$$\begin{aligned}\int_{-\infty}^{\infty} g^2(x)f(x)dx &= [g^2(x)F(x)]_{-\infty}^{\infty} - 2\int_{-\infty}^{\infty} g(x)g'(x)F(x)dx \\ &= -2\int_{-\infty}^{\infty} g(x)g'(x)F(x)dx \\ &= -2\int_{-\infty}^{\infty} \frac{g(x)F(x)}{f(x)}\sqrt{f(x)}\, g'(x)\sqrt{f(x)}\, dx.\end{aligned} \tag{16.17}$$

Applying the Cauchy–Schwartz inequality to the integral on the right side, we get that

$$\left[\int_{-\infty}^{\infty} g^2(x)f(x)dx\right]^2 \leq 4\int_{-\infty}^{\infty} \frac{g^2(x)F^2(x)}{f^2(x)} f(x)dx \int_{-\infty}^{\infty} [g'(x)]f(x)dx$$

equality occurring if and only if

$$\frac{g(x)F(x)}{f(x)} \text{ and } g'(x)$$

are linearly related. In other words

$$(E[g^2(X)])^2 \leq C_g E([g'(X)]^2) \tag{16.18}$$

where

$$C_g = 4E\left(\frac{g^2(X)F^2(X)}{f^2(X)}\right).$$

□

Remark. The inequality (16.18) is not of major interest as the constant C_g depends on the function $g(\cdot)$.

Theorem 16.6. *Suppose X is a random variable with the density function $f > 0$ on R and g is a continuously differentiable function with $E([g(X)]2) < \infty$. Further suppose that f is continuously differentiable such that*

$$\lim_{|x|\to\infty} f(x)g(x) = 0.$$

Then

$$(E[g'(X)])^2 \leq E[g^2(X)]I(f)$$

where $I(f)$ is the Fisher information.

Proof. Let $u = \sqrt{f}$ and $h = g$. Note that u and h are continuously differentiable functions such that

$$\lim_{|x|\to\infty} u^2(x)h(x) = 0.$$

Benson [1] proved that (cf. [7])

$$\int_{-\infty}^{\infty} ([u'(x)]^2 + (h'(x) + h^2(x))u^2(x))dx \geq 0.$$

Replacing h by λh and applying the above inequality, we get that

$$\lambda^2 \int_{-\infty}^{\infty} h^2(x)u^2(x)dx + \lambda \int_{-\infty}^{\infty} h'(x)u^2(x)dx + \int_{-\infty}^{\infty} [u'(x)]^2 dx \geq 0$$

for all $\lambda \in R$. Hence

$$\left[\int_{-\infty}^{\infty} h'(x)u^2(x)dx\right]^2 \leq 4\left[\int_{-\infty}^{\infty} h^2(x)u^2(x)dx\right]\left[\int_{-\infty}^{\infty} (u'(x))^2 dx\right].$$

Then it follows that

$$\left[\int_{-\infty}^{\infty} g'(x) f(x) dx\right]^2 \leq \left[\int_{-\infty}^{\infty} g^2(x) f(x) dx\right] \left[\int_{-\infty}^{\infty} \left(\frac{f'(x)}{f(x)}\right)^2 f(x) dx\right]$$

or equivalently

$$E[g'(X)])^2 \leq E[g^2(X)] I(f)$$

where $I(f)$ is the Fisher information. □

Remark. This inequality is the Cramer–Rao inequality. Other inequalities of these types are derived in [12].

16.3 Circular Random Variables

Let Z be a random variable with values on the unit circle. We identify Z with a random variable ψ such that $Z = e^{i\psi}, 0 < \psi \leq 2\pi$. For convenience in notation, we denote the typical values taken by ψ by θ. Let V be the distribution function of ψ. Then $V(0) = 0$ and $V(2\pi) = 1$. We extend the function V to R by the equation $V(\theta + 2\pi) - V(\theta) = 1$. Suppose V is absolutely continuous with the probability density function $v(\cdot)$. Then:

(a) $v(\theta) \geq 0, -\infty < \theta < \infty$
(b) $v(\theta + 2\pi) = v(\theta), -\infty < \theta < \infty$
(c) $\int_0^{2\pi} v(\theta) d\theta = 1$

(cf. [6]). Suppose there exists θ_0 such that

$$\int_0^{2\pi} (\theta - \theta_0) v(\theta) d\theta = 0. \tag{16.19}$$

Then the conditions (16.11) and (16.12) hold provided

$$\int_{\phi}^{2\pi} (\theta - \theta_0) v(\theta) d\theta \leq c v(\phi), 0 \leq \phi \leq 2\pi. \tag{16.20}$$

This can be seen by checking that

$$\int_0^{\phi} (\theta_0 - \theta) v(\theta) d\theta = \int_{\phi}^{2\pi} (\theta - \theta_0) v(\theta) d\theta \tag{16.21}$$

for $0 \leq \phi \leq 2\pi$. Hence for any absolutely continuous function $g(\cdot)$ on $[0, 2\pi]$,

$$E[g(\psi) - g(\theta_0)]^2 \leq c E([g'(\psi)]^2) \tag{16.22}$$

whenever (16.20) holds.

Suppose ψ is a circular random variable as discussed earlier with density function $v(\cdot)$. Then μ is called the *mean direction* of ψ if

$$E[\sin(\psi - \mu)] = 0$$

σ^2 is called the circular variance if

$$\sigma^2 = E[1 - \cos(\psi - \mu)] = E\left[2\sin^2\left(\frac{\psi - \mu}{2}\right)\right].$$

We now obtain an upper bound for

$$E\left[g\left(\sin\left(\frac{\psi - \theta_0}{2}\right)\right) - g(0)\right]^2$$

where $g(\cdot)$ is an absolutely continuous function. We follow the technique used in proving Theorem 16.1.

Theorem 16.7. *Suppose ψ is a circular random variable with density function v on $[0, 2\pi]$. Let*

$$\lambda(\theta) = 2\sin\left(\frac{|\theta - \theta_0|}{2}\right) \tag{16.23}$$

for $0 \le \theta, \theta_0 \le 2\pi$. Further suppose that there exists $c > 0$

$$\int_t^{2\pi} \lambda(\theta)v(\theta)d\theta \le cv(t), t > \theta_0 \tag{16.24}$$

and

$$\int_0^t \lambda(\theta)v(\theta)d\theta \le cv(t), t > \theta_0. \tag{16.25}$$

Let g be an absolutely continuous function. Then

$$E\left[g\left(\sin\left(\frac{\psi - \theta_0}{2}\right)\right) - g(0)\right]^2 \le cE\left[g'\left(\sin\left(\frac{\psi - \theta_0}{2}\right)\right)^2 \left|\cos\left(\frac{\psi - \theta_0}{2}\right)\right|\right]. \tag{16.26}$$

Proof. Note that

$$\begin{aligned} &E\left[g\left(\sin\left(\frac{\psi - \theta_0}{2}\right)\right) - g(0)\right]^2 \qquad (16.27)\\ &\le \int_0^{2\pi}\left[\int_{\theta_0}^{\theta} \frac{1}{2}g'\left(\sin\left(\frac{t - \theta_0}{2}\right)\right)\cos\left(\frac{t - \theta_0}{2}\right)dt\right]^2 v(\theta)d\theta \end{aligned}$$

$$= \int_{\theta_0}^{2\pi} \left[\int_{\theta_0}^{\theta} \frac{1}{2} g' \left(\sin \left(\frac{t-\theta_0}{2} \right) \right) \cos \left(\frac{t-\theta_0}{2} \right) dt \right]^2 v(\theta) d\theta$$

$$+ \int_{0}^{\theta_0} \left[\int_{\theta}^{\theta_0} \frac{1}{2} g' \left(\sin \left(\frac{t-\theta_0}{2} \right) \right) \cos \left(\frac{t-\theta_0}{2} \right) dt \right]^2 v(\theta) d\theta$$

$$\leq \frac{1}{4} \int_{\theta_0}^{2\pi} \left[\int_{\theta_0}^{\theta} \left| \cos \left(\frac{t-\theta_0}{2} \right) \right| dt \int_{\theta_0}^{\theta} \left[g' \left(\sin \left(\frac{t-\theta_0}{2} \right) \right) \right]^2 \left| \cos \left(\frac{t-\theta_0}{2} \right) \right| dt v(\theta) d\theta \right]$$

$$+ \frac{1}{4} \int_{0}^{\theta_0} \left[\int_{\theta}^{\theta_0} \left| \cos \left(\frac{t-\theta_0}{2} \right) \right| dt \right] \int_{\theta}^{\theta_0} \left[g' \left(\sin \left(\frac{t-\theta_0}{2} \right) \right) \right]^2 \left| \cos \left(\frac{t-\theta_0}{2} \right) \right| dt v(\theta) d\theta.$$

The proof follows from the inequalities derived in (16.27) and an application of Fubini's theorem. □

The following inequality follows as a special case from the weighted Wirtinger inequality stated in Sect. 16.3.

Theorem 16.8. *Suppose ψ is a circular random variable with a bounded, 2π-probability density function f on $[0, 2\pi]$. Further suppose that* $\inf f = L > 0$ *and* $\sup f = M < \infty$. *Then, for any absolutely continuous function $g(\cdot)$ which is 2π-periodic,*

$$E([g(\psi)]^2) \leq C(f, f) E([g'(\psi)]^2) \tag{16.28}$$

whenever $E[g(\psi)] = 0$ where

$$C(f, f) \leq \frac{\pi^2}{16} (\arctan(L^{1/2} M^{-1/2}))^{-2}. \tag{16.29}$$

Proof. Let $a = \frac{f}{L}$ and $b = \frac{f}{L}$. Then $a = b$ is bounded, 2π-periodic and $\inf_a = 1$ and $\sup_a = \frac{M}{L}$. By the weighted Wirtinger inequality, it follows that

$$\int_0^{2\pi} [g(\theta)]^2 a(\theta) d\theta \leq (a, a) \int_0^{2\pi} a(\theta [g'(\theta)]^2 d\theta$$

if g is 2π-periodic and

$$\int_0^{2\pi} g(\theta) a(\theta) d\theta = 0$$

where

$$C(a, a) \leq \frac{\pi^2}{16} (\arctan(L^{1/2} M^{-1/2}))^{-2}.$$

Rewriting these inequalities in terms of the function f, we obtain the inequality stated in (16.28). □

Remark. Ricciardi [15] proved that equality holds in (16.28) if and only if

$$\frac{f(\theta)}{L} = \bar{f}(\theta + \phi)$$

for some ϕ in R where

$$\begin{aligned} \bar{f}(\theta) = &\quad 1 \quad \text{if } 0 \le \theta < (\pi/2), \text{ and } \pi \le \theta < (3\pi/2) \\ &= M/L \text{ if } (\pi/2) \le \theta < \pi, \text{ and } (3\pi/2) \le \theta < 2\pi. \end{aligned} \tag{16.30}$$

Furthermore equality holds in (3.10) with $\frac{f(\theta)}{L} = \bar{f}(\theta + \phi)$ if and only if $g(\theta) = \bar{g}(\theta + \phi)$ where

$$\begin{aligned} \bar{g}(\theta) &= \sin\left[\sqrt{\mu}\left(\theta - \frac{\pi}{4}\right)\right] \text{if } 0 \le \theta < \frac{\pi}{2} \\ &= (M/L)^{-1/2} \cos\left[\sqrt{\mu}\left(\theta - \frac{3\pi}{4}\right)\right] \text{if } \frac{\pi}{2} \le \theta < \pi \\ &= -\sin\left[\sqrt{\mu}\left(\theta - \frac{5\pi}{4}\right)\right] \text{if } \pi \le \theta < \frac{3\pi}{2} \\ &= -(M/L)^{-1/2} \cos\left[\sqrt{\mu}\left(\theta - \frac{7\pi}{4}\right)\right] \text{if } \frac{3\pi}{2} \le \theta < 2\pi \end{aligned} \tag{16.31}$$

where $\mu = (\pi^2/16)(\arctan(M/L)^{-2})$.

It is clear that the inequality (16.28) in Theorem 16.8 can be restated in terms of the variance of the random variable $g(\psi)$ in the form

$$\text{Var}[g(\psi)] \le C(f, f)E([g'(\psi)]^2) \tag{16.32}$$

under the same conditions as stated in Theorem 16.8 without the additional condition $E[g(\psi)] = 0$.

16.4 Chernoff Type Inequality for Wrapped Normal Distribution

A circular random variable ψ is said to have the standard wrapped normal distribution if its probability density function is of the form

$$v(\theta) = (2\pi)^{1/2} \sum_{k=-\infty}^{\infty} \exp\left(-\frac{1}{2}(\theta + 2k\pi)^2\right), \quad 0 < \theta \le 2\pi. \tag{16.33}$$

This distribution is unimodal and is symmetric about $\theta = 0$. It is possible to obtain a Chernoff type inequality for the wrapped normal distribution from the inequality

(16.1) for the normal distribution by restricting the inequality to 2π-periodic functions. However we will obtain the inequality directly.

Let $g(\cdot)$ be any absolutely continuous function on $[0, 2\pi]$. Then $g(\cdot)$ is almost everywhere differentiable on $[0, 2\pi]$. Let $I(A)$ denote the indicator function of set A. Note that

$$\left(\int_0^\theta g'(\phi)d\phi\right)^2 = \left(\int_0^{2\pi} I[\theta > \phi]g'(\phi)d\phi\right)^2 \tag{16.34}$$
$$\leq \left(\int_0^{2\pi} I[\theta > \phi]d\phi\right)\left(\int_0^{2\pi} I[\theta > \phi][g'(\psi)]^2 d\phi\right)$$
$$= \theta\left(\int_0^{2\pi} I[\theta > \phi][g'(\phi)]^2 d\phi\right).$$

Hence

$$E\left[\left(\int_0^\theta g'(\phi)d\phi\right)^2\right] \leq E\left[\theta\left(\int_0^{2\pi} I[\theta > \phi][g'(\phi)]^2 d\phi\right)\right] \tag{16.35}$$
$$= \int_0^{2\pi} E(\theta I[\theta > \phi])[g'(\phi)]^2 d\phi.$$

Therefore

$$E[(g(\theta) - g(0))^2] \leq \int_0^{2\pi} E(\theta I[\theta > \phi])[g'(\phi)]^2 d\phi. \tag{16.36}$$

Observe that

$$E(\theta I[\theta > \phi]) = \frac{1}{\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} \int_\phi^{2\pi} \theta \exp\left(-\frac{1}{2}(\theta + 2k\pi)^2\right) d\theta \tag{16.37}$$
$$= \frac{1}{\sqrt{2\pi}} \int_\phi^{2\pi} \sum_{k=-\infty}^{\infty} \theta \exp\left(-\frac{1}{2}(\theta + 2k\pi)^2\right) d\theta.$$

For $k \geq 0$,

$$\int_\phi^{2\pi} \theta \exp\left(-\frac{1}{2}(\theta+2k\pi)^2\right) d\theta = \int_{\phi+2\pi}^{2k\pi+2\pi} (\alpha - 2k\pi) \exp\left(-\frac{1}{2}\alpha^2\right) d\alpha \tag{16.38}$$
$$\leq \int_{\phi+2k\pi}^{\infty} \alpha \exp\left(-\frac{1}{2}\alpha^2\right) d\alpha$$

$$= \left[-\exp\left(-\frac{1}{2}\alpha^2\right) \right]_{[\phi+2k\pi]}^{\infty}$$
$$\leq \exp\left(-\frac{1}{2}(\phi + 2k\pi)^2\right).$$

Hence

$$E(\theta I[\theta > \phi]) \leq \frac{1}{\sqrt{2\pi}} \sum_{k=-\infty}^{-1} \int_{\phi}^{2\pi} \theta \exp\left(-\frac{1}{2}(\theta + 2k\pi)^2\right) d\theta \quad (16.39)$$
$$+ \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \exp\left(-\frac{1}{2}(\phi + 2k\pi)^2\right)$$
$$= I(\phi) + \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \exp\left(-\frac{1}{2}(\theta + 2k\pi)^2\right) \text{ (say)}.$$

Combining the inequalities (16.36) and (16.39), we get that

$$E[(g(\theta) - g(0))^2] \leq \int_0^{2\pi} I(\phi)[g'(\phi)]^2 d\phi \quad (16.40)$$
$$+ \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} [g'(\phi)]^2 \sum_{k=0}^{\infty} \exp\left(-\frac{1}{2}(\phi + 2\pi)^2\right) d\phi$$
$$\leq \int_0^{2\pi} I(\phi)[g'(\phi)]^2 d\phi + E([g'(\theta)]^2).$$

16.5 Chernoff Type Inequality for von-Mises Distribution

A circular random variable ψ is said to have the von-Mises distribution if it has a probability density function of the form

$$f(\theta, \mu, k) = \frac{1}{2\pi I_0(k)} \exp[k \cos(\theta - \mu)], 0 < \theta \leq 2\pi \quad (16.41)$$

where $k > 0$ and $0 \leq \mu < 2\pi$. Here the function $I_0(k)$ is the modified Bessel function of the first kind and order zero, that is,

$$I_0(k) = \sum_{r=0}^{\infty} \frac{1}{r!^2} (k/2)^{2r}.$$

The parameter μ is called the mean direction and the parameter k is called the concentration parameter. Suppose that $\mu = 0$. We denote the normalizing constant in (16.41) by C_k and the function $f(\theta, 0, k)$ by $f_k(\theta)$ for convenience.

Let g be any absolutely continuous function. Then the derivative g' exists almost everywhere. Then

$$g(\sin\theta) - g(0) = \int_0^\theta g'(\sin u)\cos u\, du \tag{16.42}$$

for $0 \le \theta \le 2\pi$. Let

$$\begin{aligned} h(u) &= 1 \text{ if } \cos u \ge 0, \\ &= -1 \text{ if } \cos u < 0 \end{aligned} \tag{16.43}$$

for $0 \le u \le 2\pi$ and

$$\begin{aligned} p(\theta, u) &= h(u)\cos u \text{ if } 0 \le u \le \theta \\ &= 0 \qquad\qquad \text{otherwise.} \end{aligned} \tag{16.44}$$

Then

$$\begin{aligned} [g(\sin\theta) - g(0)]^2 &= \left[\int_0^\theta g'(\sin u)\cos u\, du\right]^2 \qquad (16.45) \\ &= \left[\int_0^\theta \frac{g'(\sin u)}{h(u)} h(u)\cos u\, du\right]^2 \\ &= \left[\int_0^{2\pi} \frac{g'(\sin u)}{h(u)} p(\theta, u)\, du\right]^2 \\ &\le \left[\int_0^{2\pi} p(\theta, u)\, du\right]\left[\int_0^{2\pi} \left(\frac{g'(\sin u)}{h(u)}\right)^2 p(\theta, u)\, du\right] \\ &= \left[\int_0^{2\pi} p(\theta, u)\, du\right]\left[\int_0^{2\pi} (g'(\sin u))^2 p(\theta, u)\, du\right] \\ &\quad \text{(since } h(u) \text{ is either } +1 \text{ or } -1) \\ &= \gamma(\theta)\left[\int_0^{2\pi} (g'(\sin u))^2 p(\theta, u)\, du\right] \text{(say).} \end{aligned}$$

Hence

$$\begin{aligned} E([g(\sin\psi) - g(0)]^2) &\le E\left(\gamma(\psi)\left[\int_0^{2\pi} (g'(\sin u))^2 p(\psi, u)\, du\right]\right) \qquad (16.46) \\ &= \int_0^{2\pi} E[\gamma(\psi)p(\psi, u)](g'(\sin u))^2\, du. \end{aligned}$$

Observe that

$$E[\gamma(\psi)\,p(\psi,u)] = \int_0^{2\pi} \gamma(\theta)\,p(\theta,u)\,f_k(\theta)\,d\theta \tag{16.47}$$
$$\leq \int_u^{2\pi} \gamma(\theta)h(u)\cos u\,f_k(\theta)\,d\theta$$
$$= h(u)\cos u \int_u^{2\pi} \gamma(\theta)\,f_k(\theta)\,d\theta.$$

Combining the inequalities (16.45) and (16.46), we get that

$$E([g(\sin\psi) - g(0)]^2) \leq \int_0^{2\pi} \left[h(u)\cos u \left\{\int_u^{2\pi} \gamma(\theta)\,f_k(\theta)d\theta\right\} (g'(\sin u))^2\right] du. \tag{16.48}$$

Note that

$$\int_0^{2\pi} \gamma(t)\,f_k(t)dt = \int_0^{2\pi} f_k(t)\left[\int_0^t h(u)\cos u\,d\,u\right] dt \tag{16.49}$$
$$= \int_0^{2\pi} \left[\int_0^{2\pi} f_k(t)dt\right] h(u)\cos u\,d\,u$$
$$= \int_0^{2\pi} [1 - F_k(u)]h(u)\cos u\,du$$
$$\text{(where } F_k(\theta) = \int_0^{\infty} f_k(t)dt, 0 \leq \theta \leq 2\pi)$$
$$\leq \int_0^{2\pi} |\cos u|d\,u$$
$$\leq 2\pi - \theta$$
$$\leq 2\pi.$$

The above inequalities show that

$$E([g(\sin\psi) - g(0)]^2) \leq 2\pi \int_0^{2\pi} |\cos u|(g'(\sin u))^2 du \tag{16.50}$$
$$\leq 2\pi\,E[(g'(\sin\psi))^2|\cos\psi|(f_k(\psi))^{-1}].$$

Note that $f_k(\theta) \geq C_k e^{-k}$ for all $0 \leq \theta \leq 2\pi$. Hence

$$E([g(\sin\psi) - g(0)]^2) \leq \frac{2\pi e^k}{C_k} E[(g'(\sin\psi))^2|\cos\psi|]. \tag{16.51}$$

Remark. The Chernoff type inequality for the von-Mises distribution can also be derived from the classical Wirtinger inequality by the following method. Suppose $g(\cdot)$ is a 2π-periodic absolutely continuous function with $g' \in L^2([0, 2\pi])$ and

further suppose that

$$\int_0^{2\pi} g(\theta)\, d\theta = 0.$$

Note that

$$\begin{aligned}
\mathrm{Var}[g(\psi)] &\le E[g(\psi)]^2 \qquad (16.52)\\
&\le \int_0^{2\pi} [g(\theta)]^2 f_k(\theta) d\theta\\
&\le \sup f_k(\theta) \int_0^{2\pi} [g(\theta)]^2 d\theta\\
&\le \sup f_k(\theta) \int_0^{2\pi} [g'(\theta)]^2 d\theta\\
&\le \frac{\sup f_k(\theta)}{\inf f_k(\theta)} \int_0^{2\pi} [g'(\theta)]^2 f_k(\theta) d\theta\\
&\le \frac{\sup f_k(\theta)}{\inf f_k(\theta)} E([g'(\psi)]^2)\\
&= \exp(2k) E([g'(\psi)]^2).
\end{aligned}$$

In particular,

$$\begin{aligned}
\mathrm{Var}[g(\sin(\psi))] &\le \exp(2k) E[(g'(\sin(\psi)))^2 \cos^2(\psi)] \qquad (16.53)\\
&\le \exp(2k) E[(g'(\sin(\psi)))^2 |\cos(\psi)|].
\end{aligned}$$

References

1. Benson DC (1967) Inequalities involving integrals of functions and their derivatives. J Math Anal Appl 17:292–308
2. Bobkov SG, Gotze F (1999) Exponential integrability and transportation cost related to logarithmic Sobloev inequalities. J Funct Anal 163:1–28
3. Borovkov AA, Utev SA (1983) On an inequality and a related characterization of the normal distribution. Theory Probab Appl 28:219–228
4. Brascamp HJ, Lieb EH (1976) On extensions of the Brunn–Minkowski and Prekopa–Leinder theorems, including inequalities for log concave functions and with an application to the diffusion equation. J Funct Anal 22:366–389
5. Chernoff H (1981) A note on an inequality involving the normal distribution. Ann Prob 9: 533–535
6. Mardia KV (1972) Statistics of directional data. Academic, New York
7. Mitranovic DC (1970) Analytic inequalities. Springer, Berlin
8. Muckenhoupt B (1972) Weighted norm inequalities for the Hardy maximal function. Trans Am Math Soc 165:207–226
9. Nash J (1958) Continuity of solutions of parabolic and elliptic equations. Am J Math 80: 931–954
10. Olkin I, Shepp L (2005) A matrix variance inequality. J Stat Plan Inference 130:351–358

11. Piccinini LC, Spagnolo S (1972) On the Holder continuity of second order elliptic equations in two variables. Ann Sc Norm Super Pisa 26:391–402
12. Prakasa Rao BLS (1998) Inequalities of Weyl-type and information inequalities. Indian J Pure Appl Math 29:681–694
13. Prakasa Rao BLS (2006a) Matrix variance inequalities for multivariate distributions. Stat Methodol 3:416–430
14. Prakasa Rao BLS (2006b) Some stochastic inequalities and stochastic identities. Gujarat Stat Rev 33:3–18
15. Ricciardi T (2005) A sharp weighted Wirtinger inequality, arXiv:math.AP/0501044 v1 4 Jan 2005
16. Zygmund A (1959) Trigonometric series, vol 1, 2nd rev edn. Cambridge University Press, New York, p 20

Chapter 17
Revisiting Local Asymptotic Normality (LAN) and Passing on to Local Asymptotic Mixed Normality (LAMN) and Local Asymptotic Quadratic (LAQ) Experiments

George G. Roussas and Debasis Bhattacharya

Abstract Let $X_1, \ldots, X_n$ be a random sample of size n from an underlying parametric statistical model. Then the basic statistical problem may be stated as follows: On the basis of a random sample, whose probability law depends on a parameter θ, discriminate between two values θ and θ^* ($\theta \neq \theta^*$). When the parameters are sufficiently far apart, any decent statistical procedure will do the job. A problem arises when the parameter points are close together, and yet the corresponding probability measures are substantially or even vastly different. The present paper revolves around ways of resolving such a problem. The concepts and methodology used are those of contiguity, Local Asymptotic Normality (LAN), Local Asymptotic Mixed Normality (LAMN), and Local Asymptotic Quadratic (LAQ) experiments.

17.1 Introduction, Notation, and Assumptions

A brief description of the organization of this paper is as follows. In Sect. 17.1, much of the needed notation is introduced, and assumptions under which a set of basic results, Theorems 17.1–17.7 and Theorem 17.8, are established. Some statistical applications, pertaining to testing hypotheses, are given in Sect. 17.3. In Sect. 17.5, some applications of Theorem 17.8, relative to asymptotic efficiency of estimates, are presented, whereas the next section is devoted to various comments. Generalizations of Theorems 17.1–17.7 and Theorem 17.8 are presented in Sect. 17.7. In the next section, the concepts of LAMN and LAQ are introduced, and then a set of assumptions, under which LAMN obtains, is given. In Sect. 17.9 some basic results under the LAMN framework, Theorems 17.15–17.19, are given. In Sect. 17.10, two examples, pertaining to the LAMN case, and one example pertaining to the LAQ case, are discussed, and the paper is concluded with Sect. 17.11, where the proofs of Theorems 17.1–17.6 are presented.

G.G. Roussas (✉)
Department of Statistics, University of California, Davis, CA 95616, USA
e-mail: roussas@ucdavis.edu

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_17,

Throughout the paper, all limits are taken as $n \to \infty$, unless otherwise explicitly stated. For the sake of simplicity, it will be assumed that the underlying r.v.'s are i.i.d.; in a subsequent section (see Sect. 17.7), generalizations already available will be mentioned. To this effect, let $X_0, X_1, \ldots, X_n$ be i.i.d. r.v.'s defined on the probability space $(\mathcal{X}, \mathcal{A}, P_\theta)$, $\theta \in \Theta$ open $\subseteq \mathbb{R}^k$, $k \geq 1$, let $\mathcal{A}_n = \sigma(X_0, X_1, \ldots, X_n)$ be the σ-field induced by the r.v.'s $X_0, X_1, \ldots, X_n$, and let $P_{n,\theta} = P_\theta \mid \mathcal{A}_n$ be the restriction of P_θ to $\mathcal{A}_n$. It is assumed that $P_{0,\theta}$ and P_{0,θ^*} are mutually absolutely continuous, $P_{0,\theta} \approx P_{0,\theta^*}$, $\theta, \theta^* \in \Theta$ $(\theta \neq \theta^*)$, and let

$$q(X_0; \theta, \theta^*) = \frac{\mathrm{d}P_{0,\theta^*}}{\mathrm{d}P_{0,\theta}}$$

be a specified version of the Radon–Nikodym derivative involved. Set

$$\varphi_j(\theta, \theta^*) = \varphi(X_j; \theta, \theta^*) = [q(X_j; \theta, \theta^*)]^{1/2}, \tag{17.1}$$

so that $\varphi_j(\theta, \theta^*)$ is square $P_{0,\theta}$-integrable, and the likelihood and log-likelihood functions are given, respectively, by

$$L_n(\theta, \theta^*) = \frac{\mathrm{d}P_{n,\theta^*}}{\mathrm{d}P_{n,\theta}} = \prod_{j=0}^{n} \varphi_j^2(\theta, \theta^*),$$
$$\Lambda_n(\theta, \theta^*) = \log L_n(\theta, \theta^*).$$

In the following, we restrict ourselves to θ^*, which are close to θ; i.e.,

$$\theta_n = \theta + \frac{h_n}{\sqrt{n}} \text{ with } h_n \to h \in \mathbb{R}^k. \tag{17.2}$$

Thus, the relevant likelihood and log-likelihood functions are

$$L_n(\theta, \theta_n) = \frac{\mathrm{d}P_{n,\theta_n}}{\mathrm{d}P_{n,\theta}}, \quad \Lambda_n(\theta, \theta_n) = \log L_n(\theta, \theta_n). \tag{17.3}$$

At this point, recall the definition of quadratic mean differentiability of a random function.

Definition 17.1. Let $g(\cdot; \theta)$, $\theta \in \Theta$, be a random function defined on the probability space $(\mathcal{X}, \mathcal{A}, P)$, and let $\int_{\mathcal{X}} g^2(\cdot; \theta)\, \mathrm{d}P < \infty$. Then $g(\cdot; \theta)$ is said to be *differentiable in q.m. at θ when P is employed* if

$$\frac{1}{\|h\|}[g(\cdot; \theta + h) - g(\cdot; \theta) - h'\dot{g}(\cdot; \theta)] \to 0 \text{ in q.m.}[P] \text{ as } (0 \neq)h \to 0,$$

where $\dot{g}(\cdot; \theta)$ is the derivative in q.m. of $g(\cdot; \theta)$ at θ, a k-dimensional random vector. Alternatively,

$$\frac{1}{\lambda}[g(\cdot;\theta+\lambda h)-g(\cdot;\theta)] \longrightarrow h'\dot{g}(\cdot;\theta) \text{ in q.m.}[P] \text{ as } (0<)\lambda \to 0$$

uniformly on bounded sets of values of h.

Assumptions

(A1) The probability measures $\{P_{0,\theta};\ \theta \in \Theta\}$ are mutually absolutely continuous.

(A2) (i) For each $\theta \in \Theta$, the random function $\varphi_0(\theta,\theta^*)$ (as defined in (17.1) for $j = 0$) is differentiable in q.m. with respect to θ^* at θ when P_θ is employed.
Let $\dot{\varphi}_0(\theta)$ be the derivative in q.m. of $\varphi_0(\theta,\theta^*)$ with respect to θ^* at (θ,θ) – a k-dimensional random vector. Then

(ii) $\dot{\varphi}_0(\theta)$ is $\mathcal{A}_0 \times \mathcal{C}$-measurable, where $\mathcal{C}$ is the σ-field of Borel subsets of Θ.

Let $\Gamma(\theta)$ be the covariance function defined by

$$\Gamma(\theta) = 4\mathcal{E}_\theta[\dot{\varphi}_0(\theta)\dot{\varphi}_0'(\theta)]. \tag{17.4}$$

Then

(A3) $\Gamma(\theta)$ is positive definite for every $\theta \in \Theta$.

(A4) For each $\theta \in \Theta$:

(i) $q(X_0;\theta,\theta^*) \to 1$ in $P_{0,\theta}$-probability, as $\theta^* \to \theta$.
(ii) $q(X_0;\theta,\theta^*)$ is $\mathcal{A}_0 \times \mathcal{C}$-measurable.

In terms of the q.m. derivatives $\dot{\varphi}_j(\theta)$, define the $k \times 1$ random vector $\Delta_n(\theta)$ by

$$\Delta_n(\theta) = \frac{2}{\sqrt{n}} \sum_{j=0}^{n} \dot{\varphi}_j(\theta), \tag{17.5}$$

observe that, with $\Gamma(\theta)$ defined in (17.4),

$$h'\Gamma(\theta)h = 4\mathcal{E}_\theta[h'\dot{\varphi}_0(\theta)]^2, \quad h \in \mathbb{R}^k,$$

and set

$$A(h,\theta) = \frac{1}{2}h'\Gamma(\theta)h. \tag{17.6}$$

17.2 Some Basic Results

We are now ready to state some basic results in the first part (Sects. 17.2 through 17.7) of the paper, which are the following.

Theorem 17.1. *Let θ_n, $L_n(\theta, \theta_n)$, $\Delta_n(\theta)$, and $A(h, \theta)$ be defined by* (17.2), (17.3), (17.5), *and* (17.6), *respectively. Then, under assumptions (A1)–(A4),*

$$\Lambda_n(\theta, \theta_n) - h'\Delta_n(\theta) \xrightarrow{P_{n,\theta}} -A(h, \theta).$$

Theorem 17.2. *In the same setting as that of Theorem 17.1,*

$$\mathcal{L}[\Delta_n(\theta) \mid P_{n,\theta}] \Rightarrow N(0, \Gamma(\theta)),$$

where $\Gamma(\theta)$ is defined in (17.4).

Theorem 17.3. *In the same setting as that of Theorem 17.1,*

$$\mathcal{L}[\Lambda_n(\theta, \theta_n) \mid P_{n,\theta}] \Rightarrow N\left(-\frac{1}{2}h'\Gamma(\theta)h,\ h'\Gamma(\theta)h\right).$$

An outline of the proof of these theorems will be given in Sect. 17.11. At this point, recall one mode of the definition of contiguity.

Definition 17.2. For $n \geq 0$, let P_n and Q_n be probability measures defined on the σ-field $\mathcal{A}_n$ in the measurable space $(\mathcal{X}, \mathcal{A}_n)$. Then the sequences $\{P_n\}$ and $\{Q_n\}$ are said to be *contiguous* if whenever $P_n(A_n) \to 0$, $A_n \in \mathcal{A}_n$, then $Q_n(A_n) \to 0$, and vice versa.

Remark 17.1. Although contiguity can be viewed as some sort of asymptotic mutual continuity of the pairs P_n and Q_n, it should be emphasized, however, that it is possible for $\{P_n\}$ and $\{Q_n\}$ to be contiguous without $P_n \approx Q_n$ for any n; likewise, it is possible that $P_n \approx Q_n$ for all n, but $\{P_n\}$ and $\{Q_n\}$ are not contiguous. These situations are illustrated by concrete examples. See, e.g., Examples 2.1 and 2.2, Chap. 1, in [33] (also reprinted in a paperback form in 2008).

Proposition 17.1. *Let $\{h_n^*\}$ be a bounded sequence in $\mathbb{R}^k$, and set $\theta_n^* = \theta + h_n^*/\sqrt{n}$. Then, under assumptions (A1)–(A4), the sequences $\{P_{n,\theta}\}$ and $\{P_{n,\theta_n^*}\}$ are contiguous; in particular, so are the sequences $\{P_{n,\theta}\}$ and $\{P_{n,\theta_n}\}$. An outline of the proof will be given in Sect. 17.11.*

On the basis of Theorems 17.1–17.3 and Proposition 17.1, we may establish the following results.

Theorem 17.4. *In the same setting as that of Theorem 17.1,*

$$\Lambda_n(\theta, \theta_n) - h'\Delta_n(\theta) \xrightarrow{P_{n,\theta_n}} -A(h, \theta).$$

Theorem 17.5. *In the same setting as that of Theorem 17.1,*

$$\mathcal{L}\left[\Lambda_n(\theta, \theta_n) \mid P_{n,\theta_n}\right] \Rightarrow N\left(\tfrac{1}{2}h'\Gamma(\theta)h,\ h'\Gamma(\theta)h\right).$$

Theorem 17.6. *In the same setting as that of Theorem 17.1,*

$$\mathcal{L}\left[\Delta_n(\theta) \mid P_{n,\theta_n}\right] \Rightarrow N(\Gamma(\theta)h,\ \Gamma(\theta)).$$

Remark 17.2. (i) Theorem 17.1 may be interpreted as follows: In the neighborhood of θ, for large n, and in the sense of probability,

$$\Lambda_n(\theta, \theta_n) \simeq h'\Delta_n(\theta) - A(h, \theta),$$

or by exponentiation,

$$L_n(\theta, \theta_n) \simeq \mathbb{E}^{h'\Delta_n(\theta) - A(h,\theta)}. \tag{17.7}$$

Expression (17.7) suggests that, in the neighborhood of θ and for large n, the likelihood function behaves as if it were (approximately) an exponential family. The statistical implications of this are immense. Roughly speaking, one could use the exponential family on the right-hand side of (17.7) to set up optimal tests about θ or estimates of θ, and then establish the same properties in the context of the original family of probability measures at the asymptotic level.

(ii) The precise formulation of the approximation in (17.7) is the content of Theorem 17.7 below.

(iii) Theorems 17.2 and 17.6 do demonstrate the point made earlier that, in the limit, neighboring parameter points may produce definitely distinct probability measures. The same point is made by Theorems 17.3 and 17.5.

Theorem 17.7 (Rigorous formulation of the approximation indicated in relation (17.7)). *In the same setting as that of Theorem 17.1, there exists a (suitably) truncated version $\Delta_n^*(\theta)$ of $\Delta_n(\theta)$ such that:*

$$\mathcal{E}_\theta \mathbb{E}^{h'\Delta_n^*(\theta)} \stackrel{def}{=} \mathbb{E}^{B_n(h)} < \infty,$$

$$P_{n,\theta}\left[\Delta_n^*(\theta) \neq \Delta_n(\theta)\right] \longrightarrow 0,$$
$$P_{n,\theta_n}\left[\Delta_n^*(\theta) \neq \Delta_n(\theta)\right] \longrightarrow 0,$$

and if

$$R_{n,h}(A) = \mathbb{E}^{-B_n(h)} \int_A \mathbb{E}^{h'\Delta_n^*(\theta)}\, dP_{n,\theta}, \quad A \in \mathcal{A}_n,$$

$$\left(\textit{so that } \frac{dR_{n,h}}{dP_{n,\theta}} = \mathbb{E}^{h'\Delta_n^*(\theta) - B_n(h)}, \quad h \in \mathbb{R}^k\right),$$

then

$$\|P_{n,\theta_n} - R_{n,h_n}\| \to 0, \tag{17.8}$$

or

$$\sup\left\{\left\|P_{n,\theta+\frac{h}{\sqrt{n}}} - R_{n,h}\right\|;\ h \in B \textit{ bounded } \subset \mathbb{R}^k\right\} \to 0. \tag{17.9}$$

17.3 Some Statistical Applications of Theorems 17.1–17.7

For the sake of illustration, consider first the case that Θ is a subset of $\mathbb{R}$, and consider the problem of testing a couple of hypotheses; and then consider the case that Θ is a subset of $\mathbb{R}^k$ $(k \geq 2)$ and test a simple hypothesis.

(i) $\Theta \subseteq \mathbb{R}$

Recall that in an exponential family, $f(x;\theta) = C(\theta)\mathbb{E}^{Q(\theta)T(x)}h(x)$, $x \in \mathbb{R}$, with $Q(\theta)$ being strictly monotone, there are level-α UMP tests for each one of the following hypotheses against the respective alternatives; namely,

$$H : \theta = \theta_0, \quad A : \theta > \theta_0 ; \qquad H : \theta = \theta_0, \quad A : \theta < \theta_0 ;$$

$$H : \theta \leq \theta_0, \quad A : \theta > \theta_0 ; \qquad H : \theta \geq \theta_0, \quad A : \theta < \theta_0 ;$$

$$H : \theta \leq \theta_1 \text{ or } \theta \geq \theta_2, \quad A : \theta_1 < \theta < \theta_2 \quad (\theta_1 < \theta_2).$$

(a) For hypotheses (alternatives) for which there are level-α UMP tests in the exponential family, we can construct level-α tests φ_n – based on $\Delta(\theta_j), \theta_j$, $j = 0, 1, 2$, boundary points – which are AUMP level-α tests

$$(\text{i.e., } \limsup[\sup(\mathcal{E}_\theta \omega_n - \mathcal{E}_\theta \varphi_n;\ \theta \in A)] \leq 0)$$

among all tests ω_n such that

$$\limsup\,[\sup\,(\mathcal{E}_\theta \omega_n; \theta \in H)] \leq \alpha.$$

For example, for testing $H : \theta \leq \theta_0$ against $A : \theta > \theta_0$ at level α, the test defined by:

$$\varphi_n = \varphi_n\,(\Delta_n(\theta_0)) = \begin{cases} 1, & \Delta_n(\theta_0) \geq c_n \\ \gamma_n, & \Delta_n(\theta_0) = c_n \\ 0, & \Delta_n(\theta_0) < c_n, \end{cases}$$

where c_n and γ_n are constants such that $\mathcal{E}_{\theta_0}\varphi_n = \alpha$, is an AUMP test of level α. Again, recall that in an exponential family, with $Q(\theta)$ being strictly monotone, there are level-α UMPU tests for each one of the following hypotheses against the respective alternatives; namely,
$H : \theta_1 \leq \theta \leq \theta_2$, $A : \theta < \theta_1$ or $\theta > \theta_2$ $(\theta_1 < \theta_2)$;
$H : \theta = \theta_0$, $A : \theta \neq \theta_0$.

(b) For hypotheses (alternatives) for which there are level-α UMPU tests in the exponential family, we can construct tests φ_n – based on $\Delta(\theta_j), \theta_j$, $j = 0, 1, 2$, boundary points – of asymptotic level of significance α, which are AUMP

$$(\text{i.e., } \limsup\,[\sup\,(\mathcal{E}_\theta \omega_n - \mathcal{E}_\theta \varphi_n;\ \theta \in A)] \leq 0)$$

among all tests ω_n of asymptotic level of significance α, which are asymptotically unbiased

$$(\text{i.e., } \liminf [\inf (\mathcal{E}_\theta \omega_n;\ \theta \in A)] \geq \alpha).$$

For example, for testing $H : \theta = \theta_0$ against $A : \theta \neq \theta_0$ at level α, the test φ_n defined by:

$$\varphi_n = \varphi_n(\Delta_n(\theta_0)) = \begin{cases} 1, & \Delta_n(\theta_0) < a_n \text{ or } \Delta_n(\theta_0) > b_n \\ 0, & a_n \leq \Delta_n(\theta_0) \leq b_n \end{cases},$$

$(a_n < b_n)$ with $a_n \to -\xi_{\alpha/2}$, $b_n \to \xi_{\alpha/2}$ (where ξ_p is the pth quantile of $N(0,\ \Gamma(\theta_0))$) is an AUMPU test of asymptotic level of significance α.

Remark 17.3. Actually, the tests referred to under (a) and (b) above are locally AUMP or AUMPU, respectively, but they become globally so under the additional (usually, easily fulfilled) assumption:

Assumption (A5)

$$\Delta_n(\theta_j) \overset{P_{n,\theta_n}}{\longrightarrow} \pm\infty \text{ if } \sqrt{n}(\theta_n - \theta_j) \to \pm\infty, \quad j = 0, 1, 2.$$

(ii) $\Theta \subseteq \mathbb{R}^k$, $k > 1$

For testing $H : \theta = \theta_0$ against $A : \theta \neq \theta_0$ at level α, simple tests φ_n (indicators of certain convex sets in $\mathbb{R}^k$) are constructed – based on $\Delta_n(\theta_0)$ – which enjoy Wald-type asymptotically optimal properties. These properties are too complicated to be described here in any detail, but their essence may be conveyed as follows:

(P1) The weighted average power over certain surfaces (which are ellipsoids) is largest – within a class of competing tests.

(P2) The sup of the difference of the sup and the inf of the power over certain surfaces (which are ellipsoids) $\to 0$.
At this point, recall that the *envelope power function* of a class $\mathcal{F}$ of level-α test functions φ is denoted by $\beta(\theta; \alpha)$ and is defined by $\beta(\theta; \alpha) = \sup\{\beta(\theta; \varphi);\ \varphi \in \mathcal{F}\}$, where $\beta(\theta; \varphi)$ is the power of φ at $\theta \in \Theta$. Then the next property is

(P3) The sup of the difference between the envelope power and the power over certain surfaces (which are ellipsoids) – of the test φ_n – compared to the same sup of any other competing test have a difference whose limsup is ≤ 0.

The asymptotic optimality properties of the test φ_n are implicit in the formulation of (P1)–(P3).

17.4 A Convolution Representation Theorem

In this section, we consider a class of estimates, so-called *regular* estimates, a normalized version of which is assumed to converge weakly to a probability measure.

Then this limiting probability measure assumes of a certain convolution representation.

Definition 17.3. With $\theta_n = \theta + \frac{h}{\sqrt{n}}$, the estimate T_n is said to be *regular* if

$$\mathcal{L}\left[\sqrt{n}(T_n - \theta_n) \mid P_{n,\theta_n}\right] \Rightarrow \mathcal{L}(\theta),$$

a probability measure.

Then the following result holds.

Theorem 17.8. *In the same setting as that of Theorem 17.1, and for regular estimates,*

$$\mathcal{L}(\theta) = N(0,\ \Gamma^{-1}(\theta)) * \mathcal{L}^*(\theta),$$

where $\mathcal{L}^*(\theta)$ is a specific probability measure (arising in the proof of the theorem).

Thus, the limiting measure is most concentrated when it is restricted to its normal component, and is diffused otherwise.

There are many significant applications of this theorem. Presently, we restrict ourselves to its use in the quest of asymptotically efficient estimates; this is done in the following section.

17.5 Some Applications of Theorem 17.8 in the Quest for Asymptotically Efficient Estimates

Asymptotic efficiency of an estimate is looked upon in two ways: The Weiss–Wolfowitz approach based on the asymptotic concentration of probabilities over certain classes of sets; and the classical approach based on the variance or covariance of the asymptotic distribution of the estimates.

The search for an asymptotically efficient estimate may be divided into two parts: First, establish a bound for the limit (lim sup/lim inf) of a certain desirable quantity, defined for each estimate of a target class of estimates; and second, try to identify a member of the target class for which this bound is attained. Presently, we restrict ourselves to the first part only.

17.5.1 The Weiss–Wolfowitz Approach

In this framework, there will be four results presented.

Theorem 17.9. *For $\Theta \subseteq \mathbb{R}$, consider estimates T_n such that $P_{n,\theta}\left[\sqrt{n}(T_n - \theta) \leq x\right]$ $\to F_T(x;\theta)$, a d.f.,* continuously *in Θ for each fixed $x \in \mathbb{R}$ and* continuously *on* $\mathbb{R}$

for each $\theta \in \Theta$. Let $\ell_T(\theta)$ and $u_T(\theta)$ be the "smallest" and the "largest" medians of the (continuous) d.f. $F_T(\cdot;\theta)$. Then, for every $\theta \in \Theta$,

$$\lim P_{n,\theta}\left(\theta - \frac{t_1}{\sqrt{n}} + \frac{\ell_T(\theta)}{\sqrt{n}} \le T_n \le \theta + \frac{t_2}{\sqrt{n}} + \frac{u_T(\theta)}{\sqrt{n}}\right) \le B(\theta; t_1, t_2),$$

where

$$B(\theta; t_1, t_2) = \Phi[t_2\sigma(\theta)] - \Phi[t_1\sigma(\theta)] \text{ for all } t_1, t_2 > 0,$$
$$\Phi \text{ is the d.f. of } N(0, \sigma^2(\theta)), \text{ and } \sigma^2(\theta) = 4\mathcal{E}_\theta\left[\dot{\varphi}_0(\theta)\right]^2. \tag{17.10}$$

(Recall that $\dot{\varphi}_0(\theta)$ is the q.m. derivative at θ of the square root of the p.d.f. of the underlying i.i.d. r.v.'s.)

Now, allow $\Theta \subseteq \mathbb{R}^k$, and consider estimates T_n such as those in Theorem 17.9; i.e., $P_{n,\theta}\left[\sqrt{n}(T_n - \theta) \le x\right] \to F_T(x;\theta)$, a d.f., continuously in Θ for each fixed $x \in \mathbb{R}^k$, and also continuously in $\mathbb{R}^k$ for each $\theta \in \Theta$. Then it is clear that, for each $h \in \mathbb{R}^k$, the $P_{n,\theta}\left[\sqrt{n}h'(T_n - \theta) \le x\right]$ converges to a d.f., $F_T(x;\theta,h)$, say, continuously in Θ for each fixed $x \in \mathbb{R}^k$, and also continuously in $\mathbb{R}^k$ for each fixed $\theta \in \Theta$. Furthermore, let $\ell_T(\theta,h)$ and $u_T(\theta,h)$ be the "smallest" and the "largest" medians of the (continuous) d.f. $F_T(\cdot;\theta,h)$. Then the following result holds.

Theorem 17.10. *For every $\theta \in \Theta \subseteq \mathbb{R}^k$ and each $h \in \mathbb{R}^k$, in the notation just introduced and the assumptions made, it holds*

$$\lim P_{n,\theta}\left[-t_1 + \ell_T(\theta,h) \le \sqrt{n}h'(T_n - \theta) \le t_2 + u_T(\theta,h)\right]$$
$$\le \Phi[t_2\sigma^{-1}(\theta,h)] - \Phi[-t_1\sigma^{-1}(\theta,h)],$$

for all $t_1, t_2 > 0$, where Φ is the d.f. of the $N(0, \sigma^2(\theta,h))$, and $\sigma^2(\theta,h) = h'\Gamma^{-1}(\theta)h$.

(Recall that $\Gamma^{-1}(\theta)$ is the inverse of the covariance $\Gamma(\theta)$ defined in (17.4).)

Next, for estimates T_n, where a properly normalized version converges weakly to a probability measure which has zero as its median, a result similar to that in Theorem 17.9 holds. More specifically, we have

Theorem 17.11. *Let $\Theta \subseteq \mathbb{R}$, and let T_n be an estimate such that $\mathcal{L}[\sqrt{n}(T_n - \theta)|P_{n,\theta}] \Rightarrow \mathcal{L}_{T,\theta}$, a probability measure having 0 as its median. Then*

$$\limsup P_{n,\theta}\left(\theta - \frac{t_1}{\sqrt{n}} \le T_n \le \theta + \frac{t_2}{\sqrt{n}}\right) \le B(\theta; t_1, t_2),$$

for all $t_1, t_2 > 0$, and almost all $\theta \in \Theta$ (with respect to Lebesgue measure), where $B(\theta; t_1, t_2)$ is defined in (17.10).

If one imposes median unbiasedness to the estimates considered, then a result such as that of the preceding theorem also holds. That is, we have

Theorem 17.12. *Let $\Theta \subseteq \mathbb{R}$, and let T_n be* median unbiased *estimates; i.e., for every $\theta \in \Theta$ and all n,*

$$P_{n,\theta}(T_n \geq \theta) \geq \frac{1}{2} \text{ and } P_\theta(T_n \leq \theta) \geq \frac{1}{2}.$$

Then, for all $t_1, t_2 > 0$ and all $\theta \in \Theta$, one has

$$\limsup P_{n,\theta}\left(\theta - \frac{t_1}{\sqrt{n}} \leq T_n \leq \theta + \frac{t_2}{\sqrt{n}}\right) \leq B(\theta; t_1, t_2),$$

where $B(\theta; t_1, t_2)$ is defined in (17.10)*.*

Remark 17.4. In all three cases considered above, the desirable properties of the estimates entertained are built into the definition of the classes of estimates. However, how the convolution Theorem 17.8 is entering in establishing the inequalities is not easy to explain without introducing many technical details. It suffices only to say that the concentration of probabilities under the measure $N(0, \Gamma^{-1}(\theta))$ is at least half as large as the respective concentration under the measure $\mathcal{L}(\theta)$. This is the decisive point on which the proofs hinge.

17.5.2 The Classical Approach

In this subsection, we state two results, each one for the case that $\Theta \subseteq \mathbb{R}$ and $\Theta \subseteq \mathbb{R}^k$. More precisely, we have

Theorem 17.13. *Let $\Theta \subseteq \mathbb{R}$, and let T_n be an estimate such that*

(i) Either $P_\theta\left[\sqrt{n}(T_n - \theta) \leq x\right] \to \Phi_T(x; \theta)$ for all $x \in \mathbb{R}$ and all $\theta \in \Theta$,
(ii) or $P_\theta\left[\sqrt{n}(T_n - \theta) \leq x\right] \to \Phi_T(x; \theta)$ continuously in θ for each $x \in \mathbb{R}$, where $\Phi_T(\cdot; \theta)$ is the d.f. of the $N(0, \sigma_T^2(\theta))$, for some variance $\sigma_T^2(\theta)$.

Then one has, respectively,

(i) $\sigma_T^2(\theta) \geq 1/\sigma^2(\theta)$ and $\liminf\left[n\mathcal{E}_\theta(T_n - \theta)^2\right] \geq 1/\sigma^2(\theta)$, for almost all θ (with respect to Lebesgue measure), and
(ii) $\sigma_T^2(\theta) \geq 1/\sigma^2(\theta)$ and $\liminf\left[n\mathcal{E}_\theta(T_n - \theta)^2\right] \geq 1/\sigma^2(\theta)$, for every $\theta \in \Theta$.

For the multiparameter case, there is the following generalization of parts of Theorem 17.13. Namely, we have

Theorem 17.14. *For $\theta \in \Theta \subseteq \mathbb{R}^k$, let T_n be an estimate such that*

$$P_\theta\left[\sqrt{n}(T_n - \theta) \leq x\right] \Rightarrow \Phi^{(k)}(x; C_T(\theta))$$

continuously in Θ for each $x \in \mathbb{R}^k$, where $\Phi^{(k)}(\cdot; C_T(\theta))$ is the d.f. of the $N(0, C_T(\theta))$ with $C_T(\theta)$ positive definite. Then, with $\Gamma(\theta) = \mathcal{E}_\theta\left[\dot{\varphi}_0(\theta)\dot{\varphi}_0'(\theta)\right]$,

$C_T(\theta) - \Gamma^{-1}(\theta)$ is positive semi-definite for almost all θ (with respect to Lebesgue measure). Also, $C_T(\theta)$ is continuous on Θ. Furthermore, if $\Gamma(\theta)$ is also continuous, then $C_T(\theta) - \Gamma^{-1}(\theta)$ is positive semi-definite for all $\theta \in \Theta$.

There are conditions which ensure continuity of $\Gamma(\theta)$. To this effect, make the following assumption.

Assumption (A6)

There exists $\delta > 0$ such that, for each $\theta \in \Theta$ and each $h \in \mathbb{R}^k$, there is a neighborhood of θ, $n(\theta, h)$, with the property

$$\mathcal{E}_{\theta^*} \left| h' \varphi_0(\theta^*) \right|^{2+\delta} \leq M(\theta, h) (< \infty), \text{ for all } \theta^* \in n(\theta, h).$$

Then we have the following result.

Proposition 17.2. *Under the assumptions (A1)–(A4) and (A6), $\Gamma(\theta)$ is continuous on Θ.*

To Theorem 17.14, there is the following easy but useful corollary.

Corollary 17.1. *Let g be a real-valued function defined on $\mathbb{R}^k$, and suppose that $(\partial/\partial\theta_j)[g(\theta)]$, $j = 1, \ldots, k$ exist and are continuous on Θ. Estimate $g(\theta)$ by $S_n = g(T_n)$, where the estimate T_n of θ is as in the theorem. Then one has:*

(i) $\mathcal{L}\{\sqrt{n}[S_n - g(\theta)] \mid P_{n,\theta}\} \Rightarrow \mathcal{N}(0, \sigma_S^2(\theta))$, *for all $\theta \in \Theta$, provided $\sigma_S^2(\theta) > 0$, where $\sigma_S^2(\theta) = \bar{g}'(\theta) C_T(\theta) \bar{g}(\theta)$ and*

$$\bar{g}(\theta) = \left(\frac{\partial}{\partial \theta_1} g(\theta), \ldots, \frac{\partial}{\partial \theta_k} g(\theta) \right)'.$$

(ii) $\sigma_S^2(\theta) \geq \bar{g}'(\theta) \Gamma^{-1}(\theta) \bar{g}(\theta)$ *for almost all θ (with respect to Lebesgue measure).*

(iii) *The inequality in part (ii) holds for all $\theta \in \Theta$ if $\Gamma(\theta)$ is continuous on Θ.*

17.6 Some Comments

The results stated in the preceding sections depend heavily on LAN and contiguity. These concepts have been central in much of Le Cam's research work, were introduced by him in the literature, and were given a definite form in his seminal paper [24]. They also constitute an essential part of his book [25], and they are found in a more readable version in the book [26].

The important convolution representation theorem, Theorem 17.8, was established by [17]. A version of it was also arrived at independently and at the same time by [18]. Results employing LAN and contiguity have been obtained by the first author in a series of papers. Relevant developments in a Markovian framework is

the content of the book [33,38]. Contiguity was taken a step further by [15], relating it to the statistical invariance principle.

The central assumptions among those made for the derivation of the results stated so far is the differentiability in q.m. of the square root of the p.d.f. of the underlying r.v.'s. Such an assumption may hold even if the pointwise derivative does not exist. Also, it singlehandedly replaces the Cramér-type conditions involving the existence of third-order pointwise derivatives. The interested reader may refer to the papers [27,28].

The precise formulation of the Wald-type properties referred to in Sect. 17.3(ii) may be found in Chap. 6 of the book [33,38], and, of course, in the original paper [50]. Asymptotic efficiency of estimates in terms of concentration of probability, briefly discussed in Sect. 17.5, has been studied extensively by [51–55].

17.7 Some Generalizations of Results Stated in Sects. 17.2 and 17.4

Many of the results stated in Sects. 17.2 and 17.4 have been generalized in several settings and different directions.

Thus, Theorems 17.1–17.4 have been carried over to the case when the underlying r.v.'s are independent but not identically distributed in [29]. Theorem 17.8 was discussed in a Markovian framework in [46], whereas statistical applications of Theorems 17.1–17.4 and Theorem 17.8 in the same setting were presented in [21–23]. Theorems 17.1–17.4 for Markovian processes were established originally in [31], and some statistical applications were discussed in [32,34,35]. The same theorems were carried over to the case of general discrete time-parameter stochastic processes in [36]. Theorems 17.1–17.4 were also discussed in the context of continuous time-parameter Markov processes, Lévy processes, and continuous time-parameter diffusions and Gaussian processes with known covariance by [1]. Certain aspects of the relation between contiguity, sample size, and parameter rates were addressed by [2, 3]. Theorems 17.1–17.4 were derived, under a sampling scheme based on stopping times, in [4] for Markov processes, and in [5] for general discrete time-parameter stochastic processes. The same theorems were proved in the framework of continuous time-parameter semi-Markov processes with finite state in [39]. The same problem was revisited by [41], where Theorem 17.7 was also established. In general discrete time-parameter stochastic processes, various aspects of Theorems 17.1–17.4 and Theorem 17.8 were discussed by [40,44] when stopping times are involved, and by [42,43] for fixed sample size.

17.8 The Local Asymptotic Mixed Normal and Local Asymptotic Quadratic Experiments

It must have become apparent by now that LAN is a valuable tool in discussing the asymptotic distribution of the log-likelihood functions, as well as other entities closely related to it. This is the first step in drawing statistical inference about the underlying parameter.

It so happens, however, that LAN does not obtain in some cases which are not too far out of the way. Two such cases are the explosive autoregressive process of first order, and the super-critical Galton–Watson branching process with geometric offspring distribution. What is happening instead is that, under suitable conditions, the log-likelihood function converges weakly to a Locally Asymptotically Mixed Normal (LAMN) distribution, so-termed by [19], who first arrived at this distribution.

For the definition of this concept, let δ_n be a $k \times k$ positive definite matrix such that the matrix norm of δ_n^{-1} tends to 0; i.e.,

$$\|\delta_n^{-1}\| \to 0, \tag{17.11}$$

and set

$$\theta_{nh} \text{ (to be shortened to } \theta_n) = \theta + \delta_n^{-1}h, \quad h \in \mathbb{R}^k, \tag{17.12}$$

so that $\theta_n \in \Theta$ for all sufficiently large n. Also, set

$$\Lambda_n(\theta_n, \theta) = \log \frac{\mathrm{d}P_{n,\theta_n}}{\mathrm{d}P_{n,\theta}}. \tag{17.13}$$

Then, in the language of random experiments, we have

Definition 17.4. Let δ_n, θ_n, and $\Lambda_n(\theta_n, \theta)$ be given by (17.11), (17.12), and (17.13), respectively. Then the sequence of experiments $\{(\mathcal{X}, \mathcal{A}_n, P_{n,\theta});\ \theta \in \Theta\}$, $n \geq 1$, is said to be *Locally Asymptotically Mixed Normal* (LAMN), if the following two conditions are satisfied:

(i) There exists a sequence $\{W_n(\theta)\}$, $n \geq 1$, of $\mathcal{A}_n$-measurable k-dimensional random vectors, and a sequence $\{T_n(\theta)\}$, $n \geq 1$, of $\mathcal{A}_n$-measurable $k \times k$ symmetric and a.s. $[P_{n,\theta}]$ positive definite random matrices, such that, for every $h \in \mathbb{R}^k$,

$$\Lambda_n(\theta_n, \theta) - \left[h' T_n^{1/2}(\theta) W_n(\theta) - \frac{1}{2} h' T_n(\theta) h\right] \to 0 \text{ in } P_{n,\theta}\text{-probability}. \tag{17.14}$$

(ii) There exists an a.s. $[P_{n,\theta}]$ positive definite $k \times k$ symmetric random matrix $T(\theta)$, such that,

$$\mathcal{L}\{[W_n(\theta), T_n(\theta)] | P_{n,\theta}\} \Rightarrow \mathcal{L}\{[W(\theta), T(\theta)] | P_\theta\}, \tag{17.15}$$

where $W(\theta) \sim N(0, I_k)$ and is independent of $T(\theta)$.

By setting,

$$\Delta_n(\theta) = T_n^{1/2}(\theta) W_n(\theta) \text{ and } \Delta(\theta) = T^{1/2}(\theta) W(\theta), \tag{17.16}$$

properties (i) and (ii) become as follows:

(i′) For every $h \in \mathbb{R}^k$,

$$\Lambda_n(\theta_n, \theta) - [h' \Delta_n(\theta) - \tfrac{1}{2} h' T_n(\theta) h] \to 0 \text{ in } P_{n,\theta}\text{-probability}, \tag{17.17}$$

(ii′)

$$\mathcal{L}\{[\Delta_n(\theta), T_n(\theta)] | P_{n,\theta}\} \Rightarrow \mathcal{L}\{[\Delta(\theta), T(\theta)] | P_\theta)\}. \tag{17.18}$$

If conditions (i) and (ii) (or (i′) and (ii′)) are satisfied, we also say that the sequence of families of probability measures $\{P_{n,\theta}; \theta \in \Theta\}$, $n \geq 1$, has the LAMN property at each $\theta \in \Theta$.

Remark 17.5. Comparing the convergences in Theorem 17.1 and in relation (17.17), it is seen that the constant $\Lambda(h, \theta)$ is replaced in (17.17) by the r.v. $\frac{1}{2} h' T_n(\theta) h$. Also, comparing the convergences in Theorem 17.2 and in relation (17.12), it is seen that the random vector $\Delta_n(\theta)$ is replaced in (17.18) by the random vector $[\Delta_n(\theta), T_n(\theta)]$, whose asymptotic distribution, under $P_{n,\theta}$, is $\mathcal{L}\{[\Delta(\theta), T(\theta)] \mid P_\theta\}$ rather than $N(0, \Gamma(\theta))$. See also Remark 17.7 for some additional comments.

Remark 17.6. There exist statistical experiments (to be discussed in Sect. 17.10, Example 17.3), defined by (17.17) and (17.18), but for which the relations stated in (17.16) may not hold. Such experiments are called *Locally Asymptotic Quadratic* (LAQ) experiments [20, 26]. The points at which LAN or LAMN do not hold are called 'critical points'. For LAQ experiments, it is observed that, under the contiguity of $\{P_{n,\theta}\}$ and $\{P_{n,\theta_n}\}$, the following relation holds:

$$\mathcal{E}\left[\exp(h' \Delta - \tfrac{1}{2} h' T h)\right] = 1 \quad \text{for all } h, \tag{17.19}$$

where Δ and T are as they appear in (17.18).

The assumptions made below refer to the segment of r.v.'s $X_1, \ldots, X_n$ coming from a general (discrete-time parameter) stochastic process $\{X_n\}$, $n \geq 1$; this process need not even be stationary, either in the strict or in the wide sense. In the present context, the σ-fields $\mathcal{A}_n$ are induced in $\mathcal{X}$ by the segment of r.v.'s $X_1, \ldots, X_n$.

Assumptions Pertaining to the LAMN Case

A set of assumptions will be listed below which imply the LAMN property, after some additional notation is introduced. First, it is assumed that, for $j \geq 1$, a

regular conditional probability measure of the distribution of X_j, given $\mathbf{X}_{j-1} = (X_1, \ldots, X_{j-1})$, is absolutely continuous with respect to a σ-finite measure μ_j with corresponding conditional p.d.f. $f_j(\cdot \mid \mathbf{X}_{j-1}; \theta)$, and the distribution of X_1 is absolutely continuous with respect to a σ-finite measure μ_1 with corresponding p.d.f. $f_1(\cdot; \theta)$. Furthermore, it is assumed that the support of all p.d.f.'s are independent of $\theta \in \Theta$. These assumptions imply that, for any $\theta, \theta^* \in \Theta$, $P_{n,\theta}$ and P_{n,θ^*} are mutually absolutely continuous, $P_{n,\theta} \approx P_{n,\theta^*}$, $n \geq 1$.

For notational convenience, set

$$f_j(X_j \mid \mathbf{X}_{j-1}; \theta) = f_j(\theta), \quad j \geq 2; \qquad f_1(X_1 \mid X_0; \theta) = f_1(X_1; \theta) = f_1(\theta).$$

Assumptions

(B1) With $\theta_n = \theta + \delta_n^{-1} h$ and $f_j(\theta) = f_j(X_j | \mathbf{X}_{j-1}; \theta)$, define $\xi_{nj}(\cdot; \theta, h)$ by

$$\xi_{nj}(\cdot; \theta, h) = f_j^{1/2}(\theta_n) - f_j^{1/2}(\theta) = f_j^{1/2}(\cdot \mid \mathbf{X}_{j-1}; \theta_n) - f_j^{1/2}(\cdot \mid \mathbf{X}_{j-1}; \theta).$$

Then assume that there exists a k-dimensional random vector $\xi_j(\theta)$, such that

$$\sum_{j=1}^{k} E_\theta \left\{ \int_S \left[\xi_{nj}(x_j; \theta, h) - \frac{1}{2} h' \delta_n^{-1} \xi_j(\theta) \right]^2 \mathrm{d}\mu_j \right\} \to 0,$$

(B2) Set $\eta_j(\theta) = \xi_j(\theta) / f_j^{1/2}(\theta)$, $j \geq 1$. Then assume that

$$\mathcal{E}_\theta[\eta_j(\theta) \mid \mathcal{A}_{j-1}] = 0 \quad a.s.[P_\theta], \quad j \geq 1.$$

(B3) There exists a $k \times k$ symmetric, a.s. $[P_\theta]$ positive definite random matrix $T(\theta)$, such that

$$\delta_n^{-1} \left\{ \sum_{j=1}^{n} \mathcal{E}_\theta \left[\eta_j(\theta) \eta_j'(\theta) \mid \mathcal{A}_{j-1} \right] \right\} \delta_n^{-1} - T(\theta) \to 0 \text{ in } P_\theta\text{-probability}.$$

(B4) For every $\varepsilon > 0$ and every $h \in \mathbb{R}^k$,

$$\sum_{j=1}^{n} \mathcal{E}_\theta \left\{ \left| h' \delta_n^{-1} \eta_j(\theta) \right|^2 I \left[\left| h' \delta_n^{-1} \eta_j(\theta) \right| > \varepsilon \right] \right\} \to 0,$$

where, of course, $I(\cdot)$ stands for the indicator function.

(B5) Either for every $h \in \mathbb{R}^k$, there exists $\kappa > 0$, such that

$$\sup_{n\geq 1}\sum_{j=1}^{n}\mathcal{E}_\theta\left|h'\delta_n^{-1}\eta_j(\theta)\right|^2\leq\kappa,$$

or

$$n\mathcal{E}_\theta\left|h'\delta_n^{-1}\eta_j(\theta)\right|^2=O(1),\ \text{uniformly in } j.$$

(B6) Define the r.v.'s $\eta_{nj}(\theta,h)$ by:

$$\eta_{nj}(\theta,h)=\frac{f_j^{1/2}(\theta_n)}{f_j^{1/2}(\theta)}-1.$$

Then assume that, for every $h\in\mathbb{R}^k$,

$$n\mathcal{E}_\theta\left\{[\eta_{nj}(\theta,h)+1]^4-1\right\}=O(1),\ \text{uniformly in } j\geq 1.$$

The main results to be stated in the next section also require a specification of the (fixed) matrix δ_n, the random matrices $T_n(\theta)$, and the random vectors $W_n(\theta)$. Namely, δ_n, $T_n(\theta)$, and $W_n(\theta)$ are taken to be as follows:

$$\delta_n'\delta_n=\sum_{j=1}^{n}\mathcal{E}_\theta\left[\eta_j(\theta)\eta_j'(\theta)\right],\tag{17.20}$$

$$T_n(\theta)=\delta_n^{-1}\left\{\sum_{j=1}^{n}\mathcal{E}_\theta\left[\eta_j(\theta)\eta_j'(\theta)\mid\mathcal{A}_{j-1}\right]\right\}\delta_n^{-1},\tag{17.21}$$

$$W_n(\theta)=T_n^{-1/2}(\theta)\left[\delta_n^{-1}\sum_{j=1}^{n}\eta_j(\theta)\right].\tag{17.22}$$

17.9 Some Basic Results

Here is the first result.

Theorem 17.15. *For every $h\in\mathbb{R}^k$, let θ_n be defined by (17.12), and let δ_n, $T_n(\theta)$, and $W_n(\theta)$ be defined by (17.20), (17.21), and (17.22), respectively. Then, under assumptions (B1)–(B6), the sequence of families of probability measures $\{P_{n,\theta};\theta\in\Theta\}$, $n\geq 1$, has the LAMN property; that is,*

$$\Lambda_n(\theta_n,\theta)-\left[h'T_n^{1/2}(\theta)W_n(\theta)-\tfrac{1}{2}h'T_n(\theta)h\right]\to 0\ \text{in } P_{n,\theta}\text{-probability for every } h\in\mathbb{R}^k,\tag{17.23}$$

and

$$\mathcal{L}\{[W_n(\theta), T_n(\theta)] \mid P_{n,\theta}\} \Rightarrow \mathcal{L}\{[W, T(\theta)] \mid P_\theta\}\,, \tag{17.24}$$

where $T(\theta)$ is as in assumption (B3), $W \sim N(0, I_k)$ and is independent of $T(\theta)$.

Furthermore,

Theorem 17.16. *In the setting of Theorem 17.15, it holds*

$$\mathcal{L}[\Lambda_n(\theta_n, \theta) \mid P_{n,\theta}] \Rightarrow \mathcal{L}\left[h'T^{1/2}(\theta)W - \tfrac{1}{2}h'T(\theta)h \mid P_\theta\right],$$

and therefore

$$\mathcal{L}\left[h'T_n^{1/2}(\theta)W_n(\theta) - \tfrac{1}{2}h'T_n(\theta)h \mid P_{n,\theta}\right] \Rightarrow \mathcal{L}\left[h'T^{1/2}(\theta)W - \tfrac{1}{2}h'T(\theta)h \mid P_\theta\right].$$

The following is a contiguity result. Namely,

Proposition 17.3. *In the setting of Theorem 17.15, the sequences of probability measures $\{P_{n,\theta}\}$ and $\{P_{n,\theta_n}\}$, $n \geq 1$, are contiguous.*

Theorem 17.15 and Proposition 17.3 immediately yield the following result.

Theorem 17.17. *In the setting of Theorem 17.15 and Proposition 17.3, it holds*

$$\Lambda_n(\theta_n, \theta) - \left[h'T_n^{1/2}(\theta)W_n(\theta) - \tfrac{1}{2}h'T_n(\theta)h\right] \to 0 \text{ in } P_{n,\theta_n}\text{-probability.}$$

In the following, a certain local approximation of the underlying family of probability measures is discussed. To this end, consider $\Delta_n(\theta)$ and construct a certain truncated version of it, call it $\Delta_n^*(\theta)$, like so: For $0 < k_n \uparrow \infty$, define $W_n^{k_n}$ $(= W_n^{k_n}(\theta))$ by

$$W_n^{k_n} = W_n I\left(\left|T_n^{1/2}W_n\right| \leq k_n\right) = W_n I\left(|\Delta_n| \leq k_n\right),$$

and take

$$\Delta_n^* \stackrel{def}{=} T_n^{1/2}W_n^{k_n} = T_n^{1/2}W_n I\left(|\Delta_n| \leq k_n\right) = \Delta_n I\left(|\Delta_n| \leq k_n\right).$$

Then it is seen that

$$\Delta_n^*(\theta) - \Delta_n(\theta) \to 0 \text{ in } P_{n,\theta}\text{-probability,} \tag{17.25}$$

and

$$C_{n,h}(\theta) = \int_{\mathcal{X}} \exp\left[h'\Delta_n^*(\theta) - \tfrac{1}{2}h'T_n(\theta)h\right] \mathrm{d}P_{n,\theta} < \infty. \tag{17.26}$$

Next, define the probability measures

$$Q_{n,h}(A) = C_{n,h}(\theta) \int_A \exp\left[h'\Delta_n^*(\theta) - \tfrac{1}{2}h'T_n(\theta)h\right] \mathrm{d}P_{n,\theta}, \quad A \in \mathcal{A}_n, \tag{17.27}$$

so that

$$\frac{\mathrm{d}Q_{n,h}}{\mathrm{d}P_{n,\theta}} = C_{n,h}(\theta)\exp\left[h'\Delta_n^*(\theta) - \tfrac{1}{2}h'T_n(\theta)h\right]. \tag{17.28}$$

Then the following results may be established.

Theorem 17.18. *In the notation introduced above and under assumptions (B1)–(B6):*

(a) $\sup[\|P_{n,\theta+\delta_n^{-1}h} - Q_{n,h}\|;\ h \in B$, *a bounded subset of* $\mathbb{R}^k] \to 0$.
(b) $\sup[|C_{n,h}(\theta) - 1|;\ |h| \le b] \to 0$ *for every* $b > 0$.
(c) $\|P_{n,\theta+\delta_n^{-1}h_n} - Q_{n,h_n}\| \to 0$ *for every bounded sequence* $\{h_n\}$, $n \ge 1$, *in* $\mathbb{R}^k$.

It is in the sense of Theorem 17.18 (parts (a) or (c)) that the probability measures $P_{n,\theta+\delta_n^{-1}h}$ are approximated locally (and in the sup norm mode) by the probability measures $Q_{n,h}$ defined in (17.27). Although in the construction of the measures $Q_{n,h}$ it was $\Delta_n^*(\theta)$ rather than $\Delta_n(\theta)$ which was used, this would be irrelevant asymptotically, because of (17.25). The usage of $\Delta_n^*(\theta)$ was dictated in order to ensure finiteness of the norming constant defined in (17.26).

Remark 17.7. Unlike Theorem 17.7 (see relations (17.8) and (17.9)), the approximation provided here in Theorem 17.18 is not by a standard exponential family. Instead, the approximating family is a *curved exponential family*, so termed by [14]. Roughly speaking, an exponential family is curved when the dimensionality of the sufficient statistic for θ is larger than the dimensionality of θ. For example, the normal family $N(\theta, \theta^2)$, $\theta \in \mathbb{R}$, is a curved exponential family. For more information the interested reader is referred to the book [12], and in particular Sect. 16.9.

In Theorem 17.16, the limiting distribution of the log-likelihood is a curved exponential family. However, independence of W and $T(\theta)$ and the fact that $W \sim N(0, I_k)$ turns the limiting distribution into a normal distribution, conditioned upon $T(\theta) = t$. This fact is exploited for making statistical inference in the limit, and then transposing it to the original family; optimal properties of statistical procedures will be of the asymptotic variety. Relevant references in this respect are [6, 7, 13, 16, 47–49]. The proofs of Theorems 17.15–17.18, when stopping times are involved, can be found in [8–11]. A thorough discussion of this sort of asymptotic results with applications to time series models can be found in [20].

A convolution representation result similar to that in Theorem 17.8 is established here, too. More precisely, one has

Theorem 17.19. *Suppose assumptions (B1)–(B6) hold. Also, let* $\{V_n\}$, $n \ge 1$, *be a sequence of* k*-dimensional random vectors of regular estimates of* θ*; that is, for every* $h \in \mathbb{R}^k$, *there is a* k*-dimensional random vector* $V(\theta)$, *such that*

$$\mathcal{L}\{[\delta_n(V_n - \theta_n), T_n(\theta)] \mid P_{n,\theta_n}\} \Rightarrow \mathcal{L}\{[V(\theta), T(\theta)] \mid P_\theta\} = L(\theta),\ \textit{say}. \tag{17.29}$$

Let $L_{T(\theta)}$ *be a regular version of the conditional distribution of* $V(\theta)$, *given* $T(\theta)$. *Then*

$$L_{V(\theta)|T(\theta)} = L_{T(\theta)} = L_1 * L_2 \quad a.s.\ [P_\theta], \tag{17.30}$$

where $L_1 = N(0, T^{-1}(\theta))$ *and* L_2 *is the conditional distribution (under* P_θ*) of* $V(\theta) - T^{-1/2}(\theta)Z(\theta)$*, given* $T(\theta)$*, and* $Z(\theta)$ *follows the* $N(0, I_k)$ *distribution and is independent of* $T(\theta)$*.*

It is to be noted that the convolution representation here refers to the conditional distribution of $V(\theta)$, given $T(\theta)$, unlike the case in Theorem 17.8. This is the way one may circumvent the curvature of the limiting family in the present context. A version of Theorem 17.19, when stopping times are involved, can be found in [45].

17.10 Examples Pertaining to the LAMN and LAQ Cases

Here are three examples whose discussion falls into the LAMN and LAQ frameworks.

Example 17.1 (Explosive autoregressive process of first order). Such a process consists of r.v.'s X_j, $j \geq 0$, generated as follows:

$$X_j = \theta X_{j-1} + \varepsilon_j, \quad X_0 = 0, \quad |\theta| > 1,$$

where the ε_j's are independent r.v.'s distributed as $N(0, 1)$. The X_j's form a Markov process with transition p.d.f. that of $N(\theta x_{j-1}, 1)$, so that

$$f_j(x_j|\mathbf{x}_{j-1}; \theta) = f(x_j|x_{j-1}; \theta) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x_j - \theta x_{j-1})^2\right].$$

In [8], it has been checked that assumptions (B1)–(B6) hold, so that the underlying family of probability measures is LAMN (see also [16, p. 110], and [49, pp. 135–136]).

The key quantities here are:

$$\delta_n^{-1} = \frac{\theta^2 - 1}{\theta^n}, \quad \text{so that } \theta_n = \theta + \frac{(\theta^2 - 1)h}{\theta^n},$$

$$\xi_j(\theta) = \frac{\dot{f}_j(\theta)}{f_j^{1/2}(\theta)},$$

where

$$\dot{f}_j(\theta) = \frac{\partial}{\partial\theta} f(x_j|x_{j-1}; \theta) = \frac{x_{j-1}(x_j - \theta x_{j-1})}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x_j - \theta x_{j-1})^2\right],$$

so that

$$\eta_j(\theta) = \frac{\xi_j(\theta)}{f_j^{1/2}(\theta)} = \frac{\dot{f}_j(\theta)}{f_j(\theta)} = x_{j-1}(x_j - \theta x_{j-1}).$$

It follows that

$$T_n(\theta) = \frac{(\theta^2-1)^2}{\theta^{2n}} \sum_{j=1}^{n} X_{j-1}^2, \quad W_n(\theta) = \frac{\sum_{j=1}^{n} X_{j-1}\epsilon_j}{\left(\sum_{j=1}^{n} X_{j-1}^2\right)^{1/2}},$$

$$\Delta_n(\theta) = \frac{(\theta^2-1)}{\theta^n} \sum_{j=1}^{n} X_{j-1}\epsilon_j.$$

Furthermore, it is seen that the log-likelihood is given by

$$\Lambda_n(\theta_n, \theta) = \frac{(\theta^2-1)h}{\theta^n} \frac{\sum_{j=1}^{n} X_{j-1}(X_j - \theta X_{j-1})}{(\sum_{j=1}^{n} X_{j-1}^2)^{1/2}} - \frac{(\theta^2-1)^2h^2}{2\theta^{2n}} \sum_{j=1}^{n} X_{j-1}^2 + o_{P_{n,\theta}}(1).$$

From results found in [6, pp. 164–165] and [16, p. 110], it is seen that

$$\mathcal{L}[T_n(\theta)|P_{n,\theta}] \Rightarrow \mathcal{L}[T(\theta)|P_\theta] = \chi_1^2,$$
$$\mathcal{L}\{[\Delta_n(\theta), T_n(\theta)]|P_{n,\theta}\} \Rightarrow \mathcal{L}\{[T^{1/2}(\theta)W, T(\theta)]|P_\theta\},$$

where $W \sim N(0,1)$, $T^{1/2}(\theta) \sim N(0,1)$, and $T^{1/2}(\theta)$ and W are independent (all under P_θ).

Example 17.2 (Super-critical Galton–Watson branching process with geometric offspring distribution). This example has been discussed in [6, p. 163]; [13, pp. 853–854]; [8]; and [49, pp. 133–135]. In the present case,

$$f_j(x_j|\mathbf{x}_{j-1};\theta) = f(x_j|x_{j-1};\theta) = \left(1 - \frac{1}{\theta}\right)^{x_j - x_{j-1}} \left(\frac{1}{\theta}\right)^{x_{j-1}}, \quad \theta > 1.$$

Here, the geometric offspring distribution is given by

$$P(X_1 = j) = \theta^{-1}(1-\theta^{-1})^{j-1}, \quad j = 1, 2, \ldots, \quad \theta > 1,$$

so that

$$E_\theta X_1 = \theta \text{ and } \mathrm{Var}_\theta(X_1) = \theta(\theta-1).$$

The key quantities here are (see [13, pp. 853–854] and [49, pp. 133–135]):

$$\delta_n^{-1} = \frac{\theta^{1/2}(\theta-1)}{\theta^{n/2}}, \text{ so that } \theta_n = \theta + \frac{\theta^{1/2}(\theta-1)h}{\theta^{n/2}},$$

$$\xi_j(\theta) = \frac{\dot{f}_j(\theta)}{f_j^{1/2}(\theta)}, \text{ where } \frac{\partial}{\partial\theta}\log f(x_j|x_{j-1};\theta) = \frac{\partial}{\partial\theta}\log f_j(\theta) = \frac{x_j - \theta x_{j-1}}{\theta(\theta-1)},$$

$$\text{and } \dot{f}_j(\theta) = f_j(\theta) \times \frac{\partial}{\partial\theta}\log f_j(\theta), \text{ so that } \eta_j(\theta) = \frac{X_j - \theta X_{j-1}}{\theta(\theta-1)}.$$

It follows that

$$T_n(\theta) = \frac{\theta-1}{\theta^n}\sum_{j=1}^{n} X_{j-1}, \quad W_n(\theta) = \frac{\sum_{j=1}^{n}\left(X_j - \theta X_{j-1}\right)}{[\theta(\theta-1)\sum_{j=1}^{n} X_{j-1}]^{1/2}},$$

and

$$\Delta_n(\theta) = \frac{\sum_{j=1}^{n}\left(X_j - \theta X_{j-1}\right)}{\theta^{(n-1)/2}}.$$

Furthermore, the log-likelihood function is given by

$$\Lambda_n(\theta_n,\theta) = \frac{h}{\theta^{(n-1)/2}}\sum_{j=1}^{n}(X_j - \theta X_{j-1}) - \frac{\theta(\theta-1)^2h^2}{2\theta^n}\sum_{j=1}^{n} X_{j-1} + o_{P_{n,\theta}}(1).$$

It can be seen (from the references just cited) that

$$\begin{aligned}
\mathcal{L}[T_n(\theta)|P_{n,\theta}] &\Rightarrow \mathcal{L}[T(\theta)],\\
\mathcal{L}\{[\Delta_n(\theta), T_n(\theta)]|P_{n,\theta}\} &\Rightarrow \mathcal{L}\{[T^{1/2}(\theta)W, T(\theta)]|P_\theta\},\\
\mathcal{L}[\Delta_n(\theta)|P_{n,\theta}] &\Rightarrow \mathcal{L}[T^{1/2}(\theta)W|P_\theta],
\end{aligned}$$

where $T(\theta)$ is an exponentially distributed random variable (under P_θ) with unit mean, $W \sim N(0, I_k)$, and $T^{1/2}(\theta)$ and W are independent.

Example 17.3. Let us consider the process described in Example 17.1 again, where $\theta = 1$ is the true value of the parameter θ (a unit root autoregressive process of order one). We use (17.20) to find δ_n as indicated below, and, without loss of generality, take $\delta_n = n$, so that $\theta_n = \theta + \frac{h}{n}$.

$$\delta_n^2 = \mathcal{E}\left(\sum_{j=1}^{n} X_{j-1}^2\right) = \frac{n(n-1)}{2} \simeq n^2.$$

We find that the log-likelihood function is given by

$$\Lambda_n(\theta_n,\theta) = \frac{h}{n}\sum_{j=1}^{n} X_{j-1}\varepsilon_j - \frac{h^2}{2n^2}\sum_{j=1}^{n} X_{j-1}^2 + o_{P_{n,\theta}}(1).$$

It can be seen that the distribution of $\Lambda_n(\theta_n,\theta)$ tends weakly to the limit distribution of Λ_h, where

$$\Lambda_h = \int_0^1 W_t \, \mathrm{d}W_t - \frac{h^2}{2} \int_0^1 W_t^2 \, \mathrm{d}t = \frac{1}{2}(W_1^2 - 1) - \frac{h^2}{2} \int_0^1 W_t^2 \, \mathrm{d}t,$$

and W_t is a Weiner process or standard Brownian motion [16, 30, 48]. Here

$$\Delta_n(\theta) = \frac{1}{n} \sum_{j=1}^{n} X_{j-1} \varepsilon_j \Rightarrow \int_0^1 W_t \, \mathrm{d}W_t = \frac{1}{2}(W_1^2 - 1),$$

where W_t^2 is a $N(0, t)$ r.v. for fixed t,

$$T_n(\theta) = \frac{1}{n^2} \sum_{j=1}^{n} X_{j-1}^2 \Rightarrow \int_0^1 W_t^2 \, \mathrm{d}t,$$

and

$$\mathcal{L}\{[\Delta_n(\theta), T_n(\theta)] \mid P_{n,\theta}\} \Rightarrow \mathcal{L}[(\Delta, T) \mid P_\theta] = \left(\int_0^1 W_t \, \mathrm{d}W_t, \int_0^1 W_t^2 \, \mathrm{d}t \right).$$

It can also be seen that

$$\mathcal{L}[n(\widehat{\theta}_n - 1) \mid P_{n,\theta}] \Rightarrow \mathcal{L}\left[\tfrac{1}{2}(W_1^2 - 1) \Big/ \textstyle\int_0^1 W_t^2 \, \mathrm{d}t \mid P_\theta \right],$$

where

$$\widehat{\theta}_n = \frac{\sum_{j=1}^n X_j X_{j-1}}{\sum_{j=1}^n X_{j-1}^2}$$

and

$$n(\widehat{\theta}_n - 1) = \frac{\frac{1}{n} \sum_{j=1}^n X_{j-1}(X_j - X_{j-1})}{\frac{1}{n^2} \sum_{j=1}^n X_{j-1}^2} = \frac{\frac{1}{n} \sum_{j=1}^n X_{j-1} \varepsilon_j}{\frac{1}{n^2} \sum_{j=1}^n X_{j-1}^2} = \frac{\Delta_n(\theta)}{T_n(\theta)}.$$

17.11 Outline of Proofs of Some of the Basic Results

From Sect. 17.1, recall that $\varphi_j(\theta, \theta^*) = [q(X_j; \theta, \theta^*)]^{1/2}$. Keep θ fixed, replace θ^* by $\theta_n = \theta + h_n/\sqrt{n}$ with $h_n \to h \in \mathbb{R}^k$, and set

$$\varphi_{nj}(\theta) = \varphi_j(\theta, \theta_n) = [q(X_j; \theta, \theta_n)]^{1/2}.$$

Also, recall that $\dot{\varphi}_0(\theta)$ is the q.m. derivative of $\varphi_0(\theta, \theta^*)$ with respect to θ^* at θ when P_θ is employed. In this notation and always under assumptions (A1)–(A4), we have

Lemma 17.1.

$$\sum_{j=0}^{n}[\varphi_{nj}(\theta)-1]^2 \to \mathcal{E}_\theta[h'\dot{\varphi}_0(\theta)]^2 \text{ in } P_\theta\text{-probability.}$$

Proof. See Lemma 5.1 in [33, 38]. □

Lemma 17.2. *With* $\Lambda_n(\theta,\theta_n)=\Lambda_n(\theta)$ *defined by* (17.3)*, we have*

$$\Lambda_n(\theta)-2\sum_{j=0}^{n}[\varphi_{nj}(\theta)-1]+\sum_{j=0}^{n}[\varphi_{nj}(\theta)-1]^2 \to 0 \text{ in } P_\theta\text{-probability.}$$

Proof. See Lemma 5.3 in [33, 38]. □

From Lemmas 17.1 and 17.2 we have, by subtraction

$$\Lambda_n(\theta)-2\sum_{j=0}^{n}[\varphi_{nj}(\theta)-1] \to -\mathcal{E}_\theta[h'\dot{\varphi}_0(\theta)]^2 \text{ in } P_\theta\text{-probability.} \tag{17.31}$$

Now, in the alternative definition of q.m. differentiability (see Definition 17.1), replace θ^* by $\theta+\frac{h}{\sqrt{n}}$ to obtain

$$\sqrt{n}\left[\varphi_j\left(\theta,\theta+\frac{h}{\sqrt{n}}\right)-1\right] \underset{(P_\theta)}{\overset{q.m.}{\longrightarrow}} h'\dot{\varphi}_j(\theta),$$

since $\varphi_j(\theta,\theta)=1$; or

$$\sqrt{n}\left[\varphi_{nj}(\theta)-1\right] \underset{(P_\theta)}{\overset{q.m.}{\longrightarrow}} h'\dot{\varphi}_j(\theta). \tag{17.32}$$

However,

Lemma 17.3. $\sqrt{n}[\varphi_{nj}^2(\theta)-1] \to 2h'\dot{\varphi}_j(\theta)$ *in the first mean* $[P_\theta]$.

Proof. See Lemma 5.4 in [33, 38]. □

But if $U_n \overset{q.m.}{\longrightarrow} U$, then

$$|\mathcal{E}U_n-\mathcal{E}U| = |\mathcal{E}(U_n-U)| \le \mathcal{E}|U_n-U| \le \mathcal{E}^{1/2}|U_n-U|^2 \to 0$$

(see, e.g., Corollary to Theorem 2, Chapter 6, in [37]). Accordingly,

$$\mathcal{E}_\theta\left[\sqrt{n}\left(\varphi_{nj}^2(\theta)-1\right)\right] \longrightarrow \mathcal{E}_\theta[2h'\dot{\varphi}_j(\theta)].$$

However, $\mathcal{E}_\theta\varphi_{nj}^2(\theta)=1$, so that $\mathcal{E}_\theta[h'\dot{\varphi}_j(\theta)]=0$; Thus, we have the following lemma.

Lemma 17.4. *For each* $h \in \mathbb{R}^k$, $\mathcal{E}_\theta[h'\dot{\varphi}_j(\theta)] = 0$.

Next, omit θ in the expressions for $\varphi_{nj}(\theta)$ and $\dot{\varphi}_j(\theta)$, and set

$$Y_j = (\varphi_{nj} - 1) - \frac{1}{\sqrt{n}} h'\dot{\varphi}_j - (\mathcal{E}_\theta \varphi_{nj} - 1),$$

so that $\mathcal{E}_\theta Y_j = 0$, and apply Tchebichev's inequality to get the result.

$$P_\theta\left(\left|\sum_{j=0}^n Y_j\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2}\sigma_\theta^2\left(\sum_{j=0}^n Y_j\right) = \frac{n+1}{\varepsilon^2}\sigma_\theta^2\left[(\varphi_{n0} - 1) - \frac{1}{\sqrt{n}} h'\dot{\varphi}_0\right]$$
$$\leq \frac{n+1}{\varepsilon^2}\mathcal{E}_\theta\left|\sqrt{n}\,(\varphi_{n0} - 1) - h'\dot{\varphi}_0\right|^2 \longrightarrow 0,$$

so that

$$\sum_{j=0}^n Y_j = \sum_{j=0}^n (\varphi_{nj} - 1) - \frac{1}{\sqrt{n}}\sum_{j=0}^n h'\dot{\varphi}_j - (n+1)(\mathcal{E}_\theta\varphi_{n0} - 1) \to 0 \text{ in } P_\theta\text{-probability}. \tag{17.33}$$

Combining (17.31) and (17.33), we get

$$\Lambda_n - \frac{2}{\sqrt{n}}\sum_{j=0}^n h'\dot{\varphi}_j - 2(n+1)(\mathcal{E}_\theta\varphi_{n0} - 1) \to -\mathcal{E}_\theta\left(h'\dot{\varphi}_0\right)^2 \text{ in } P_\theta\text{-probability}. \tag{17.34}$$

From the identity

$$\varphi_{n0}^2 - 1 = (\varphi_{n0} - 1)^2 + 2(\varphi_{n0} - 1),$$

we obtain, by taking expectations,

$$0 = \mathcal{E}_\theta\left(\varphi_{n0}^2 - 1\right) = \mathcal{E}_\theta(\varphi_{n0} - 1)^2 + 2(\mathcal{E}_\theta\varphi_{n0} - 1), \text{ or}$$

$$\mathcal{E}_\theta\left[\sqrt{n}\,(\varphi_{n0} - 1)\right]^2 + 2n(\mathcal{E}_\theta\varphi_{n0} - 1) = 0. \tag{17.35}$$

Since $\mathcal{E}_\theta\left[\sqrt{n}\,(\varphi_{n0} - 1)\right]^2 \to \mathcal{E}_\theta(h'\dot{\varphi}_0)^2$, relation (17.35) yields

$$2n(\mathcal{E}_\theta\varphi_{n0} - 1) \to -\mathcal{E}_\theta\left(h'\dot{\varphi}_0\right)^2,$$

or

$$2(n+1)(\mathcal{E}_\theta\varphi_{n0} - 1) \to -\mathcal{E}_\theta\left(h'\dot{\varphi}_0\right)^2. \tag{17.36}$$

Relations (17.34) and (17.36) yield then

$$\Lambda_n - \frac{2}{\sqrt{n}}\sum_{j=0}^n h'\dot{\varphi}_j \longrightarrow -2\mathcal{E}_\theta\left(h'\dot{\varphi}_0\right)^2 \text{ in } P_\theta\text{-probability}. \tag{17.37}$$

Since, by relations (17.6) and (17.4),

$$-A(h,\theta) = -\frac{1}{2}h'\Gamma(\theta)h = -\frac{1}{2}\times 4\mathcal{E}_\theta\left[h'\dot{\varphi}_0(\theta)\right]^2 = -2\mathcal{E}_\theta\left[h'\dot{\varphi}_0(\theta)\right]^2,$$

and

$$\frac{2}{\sqrt{n}}\sum_{j=0}^{n}\dot{\varphi}_j(\theta) = \Delta_n(\theta),$$

by (17.5), then relation (17.37) establishes Theorem 17.1; i.e.,

Proof of Theorem 17.1.

$$\Lambda_n(\theta) - h'\Delta_n(\theta) \longrightarrow -A(h,\theta) \text{ in } P_{n,\theta}\text{-probability.}$$

□

We proceed now with the proof of Theorem 17.2.

Proof of Theorem 17.2. By dropping θ for convenience, it suffices to show that, for all $h \in \mathbb{R}^k$,

$$\mathcal{L}\left(h'\Delta_n \mid P_{n,\theta}\right) \Rightarrow N(0,\ h'\Gamma h).$$

From (17.5),

$$h'\Delta_n = \frac{2}{\sqrt{n}}\sum_{j=0}^{n}h'\dot{\varphi}_j$$

with the r.v.'s $h'\dot{\varphi}_j$, $j = 0, 1, \ldots, n$, being i.i.d., $\mathcal{E}_\theta\left(h'\dot{\varphi}_0\right) = 0$ and $\sigma_\theta^2\left(h'\dot{\varphi}_0\right) < \infty$. Then the CLT applies and gives

$$\frac{1}{\sqrt{n}}\sum_{j=0}^{n}h'\dot{\varphi}_j \Rightarrow N\left(0,\ \mathcal{E}_\theta\left(h'\dot{\varphi}_0\right)^2\right) \text{ under } P_\theta,$$

or

$$\mathcal{L}[h'\Delta_n(\theta) \mid P_{n,\theta}] \Rightarrow N(0,\ h'\Gamma(\theta)h) \text{ for all } h \in \mathbb{R}^k,$$

so that

$$\mathcal{L}[\Delta_n(\theta) \mid P_{n,\theta}] \Rightarrow N(0,\ \Gamma(\theta)).$$

□

Proof of Theorem 17.3. It is an immediate consequence of Theorems 17.1 and 17.2.

□

At this point, the contiguity result stated in Proposition 17.1 is established.

Proof of Proposition 17.1. By Theorem 17.3,

$$\mathcal{L}\left(\Lambda_n \mid P_{n,\theta}\right) \Rightarrow N\left(-\frac{\sigma^2}{2},\ \sigma^2\right),$$

where $\sigma^2 = h'\Gamma(\theta)h$, and

$$\int_{\mathbb{R}} \mathbb{E}^{\lambda}\, \mathrm{d}N\left(-\frac{\sigma^2}{2},\ \sigma^2\right) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma}\mathbb{E}^{\lambda} \exp\left\{-\left(\lambda + \frac{\sigma^2}{2}\right)^2 /2\sigma^2\right\}\ \mathrm{d}\lambda$$
$$= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\left(\lambda - \frac{\sigma^2}{2}\right)^2 /2\sigma^2\right\}\ \mathrm{d}\lambda = 1.$$

Then statement (S_3) on page 11 in [33, 38] implies statement (S_1), which is the required contiguity. □

Proof of Theorem 17.4. Contiguity of $\{P_{n,\theta}\}$ and $\{P_{n,\theta_n}\}$ implies that, for any $\mathcal{A}_n$-measurable r.v.'s T_n, the convergence $T_n \to 0$ in $P_{n,\theta}$-probability implies $T_n \to 0$ in P_{n,θ_n}-probability, and vice versa. This is so by Proposition 2.1, page 8, in [33,38]. Applying this for $\Lambda_n(\theta) = \Lambda_n(\theta, \theta_n)$, we obtain the desired result. □

Proof of Theorem 17.5. It follows by Theorem 17.3, the contiguity of the sequences $\{P_{n,\theta}\}$ and $\{P_{n,\theta_n}\}$, and Corollary 7.2, page 35 in [33, 38]. □

Proof of Theorem 17.6. It follows by Theorem 7.2, page 38 in [33, 38], by way of Theorems 17.1, 17.2, and Proposition 17.1. □

The remaining main results are considerably more complicated, and any attempt to provide even an outline of their proofs would take us too far afield. So, this section is concluded here. Interested readers are referred to the different papers written by the authors for the omitted proofs of results stated herein.

References

1. Akritas MG (1978) Contiguity of probability measures associated with continuous time stochastic processes. PhD thesis, Department of statistics, University of Wisconsin, Madison
2. Akritas MG, Puri ML, Roussas GG (1979) Sample size, parameter rates and contiguity–The iid case. Commun Stat Theory Methods 8(1):71–83
3. Akritas MG, Roussas GG (1978) A characterization of contiguity by sample size and parameter rates. Symposia Mathematica, Instituto Nazionale di Alta Mathematica, vol 25. Academic Press, New York, pp 155–170
4. Akritas MG, Roussas GG (1979) Asymptotic expansion of the log-likelihood function based on stopping times defined on a Markov process. Ann Inst Stat Math 31(Part A):103–120
5. Akritas MG, Roussas GG, Stamatelos GD (1984) Asymptotic expansion of the log-likelihood function based on stopping times defined on stochastic processes. Mathematical structures – Computational mathematics – Mathematical modelling, Bulgarian Academy of Sciences, Sofia, 90–96
6. Basawa IV, Brockwell PJ (1984) Asymptotic conditional inference for regular nonergodic models with an application to autoregressive processes. Ann Stat 12:161–171
7. Basawa IV, Scott DJ (1983) Asymptotic optimal inference for non-ergodic models. Springer, New York
8. Basu AK, Bhattacharya D (1988) Local asymptotic mixed normality of log-likelihood based on stopping times. Calcutta Stat Assoc Bull 37:143–159
9. Basu AK, Bhattacharya D (1990) Weak convergence of randomly stopped log-likelihood ratio statistics to mixed Gaussian process. Calcutta Stat Assoc Bull 39:137–149

10. Basu AK, Bhattacharya D (1992) On the asymptotic non-null distribution of randomly stopped log-likelihood ratio statistic. Calcutta Stat Assoc Bull 42:255–260
11. Bhattacharya D, Roussas GG (2001) Exponential approximation for randomly stopped locally asymptotically mixture of normal experiments. Stoch Model Appl 4(2):56–71
12. DasGupta A (2008) Asymptotic theory of statistics and probability. Springer, New York
13. Davies RB (1985) Asymptotic inference when the amount of information is random. In: Le Cam L, Olson RA (eds) Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, vol 2. Wadsworth, California, pp 841–864
14. Efron B (1975) Defining the curvature of a statistical problem (with applications to second order efficiency). Ann Stat 3(6):1189–1242
15. Greenwood PE, Shiryayev AN (1985) Contiguity and the statistical invariance principle. Gordon and Breach, New York
16. Greenwood PE, Wefelmeyer W (1993) Asymptotic minimax results for stochastic process families with critical points. Stoch Process Appl 44:107–116
17. Hájek JA (1970) A characterization of limiting distributions of regular estimates. Z Wahrschelnlichkeitstheoric Und verw Gebiete 14:323–330
18. Inagaki N (1970) On the limiting distribution of sequence of estimators with uniformity property. Ann Inst Stat Math 22:1–13
19. Jeganathan P (1982) On the asymptotic theory of estimation when the limit of the log-likelihood ratios is mixed normal. Sankhyā 44(Ser. A):173–212
20. Jeganathan P (1995) Some aspects of asymptotic theory with applications to time series model. Econom Theory 2:818–887
21. Johnson RA, Roussas GG (1969) Asymptotically most powerful tests in Markov processes. Ann Math Stat 40:1207–1215
22. Johnson RA, Roussas GG (1970) Asymptotically optimal tests in Markov processes. Ann Math Stat 41:918–38
23. Johnson RA, Roussas GG (1972) Applications of contiguity to multiparameter hypothesis testing. Proceedings of the sixth Berkeley symposium on probability theory and mathematical statistics, vol 1, pp 195–226
24. Le Cam L (1960) Locally asymptotically normal families of distributions. Univ Calif Publ Stat 3:37–98
25. Le Cam L (1986) Asymptotic methods in statistical decision theory. Springer series in statistics. Springer, New York
26. Le Cam L, Yang GL (2000) Asymptotics in statistics, some basic concepts. Lecture notes in statistics. Springer, New York
27. Lind B, Roussas GG (1972) A remark on quadratic mean differentiability. Ann Math Stat 43:1030–1034
28. Lind B, Roussas GG (1977) Cramér-type conditions and quadratic mean differentiability. Ann Inst Stat Math 29:189–201
29. Philippou A, Roussas GG (1973) Asymptotic distribution of the likelihood function in the independent not identically distributed case. Ann Stat 1:454–471
30. Phillips PCB (1987) Asymptotic minimax results for stochastic process families with critical points. Econometrica 55(2):277–301
31. Roussas GG (1965) Asymptotic inference in Markov processes. Ann Math Stat 36:978–993
32. Roussas GG (1968) Some applications of the asymptotic distribution of likelihood functions to the asymptotic efficiency of estimates. Z Wahrscheinlichkeitstheorie Und verw Gebeite 10:252–260
33. Roussas GG (1972) Contiguity of probability measures: Some applications in statistics. Cambridge University Press, Cambridge
34. Roussas GG (1975) Asymptotic properties of maximum probability estimates in the iid case. In: Puri ML (ed) Statistical inference and related topics, vol 2. Academic Press, New York, pp 211–234
35. Roussas GG (1977) Asymptotic properties of maximum probability estimates in Markov processes. Ann Inst Stat Math 29:203–219

36. Roussas GG (1979) Asymptotic distribution of the log-likelihood function for stochastic processes. Z Wahrschelnlichkeitstheoric Und verw Gebiete 47:31–46
37. Roussas GG (2005) An introduction to measure-theoretic probability. Elsevier, Burlington
38. Roussas GG (2008) Contiguity of probability measures: Some applications in statistics. Cambridge University Press, paperback reprint of the book #33
39. Roussas GG, Akritas MG (1980) Asymptotic inference in continuous time semi-Markov processes. Scand J Stat 7:73–79
40. Roussas GG, Bhattacharya D (1999a) Asymptotic behavior of the log-likelihood function in stochastic processes when based on a random number of random variables. In: Janssen J, Limnios N (eds) Semi-Markov models and applications. Kluwer, Dordrecht, pp 119–147
41. Roussas GG, Bhattacharya D (1999b) Some asymptotic results and exponential approximation in semi-Markov processes. In: Janssen J, Limnios N (eds) Semi-Markov models and applications, pp 149–166
42. Roussas GG, Bhattacharya D (2002) Exponential approximation of distributions. In Teoriya Imovirnostey ta Matematichna Statystika, 66:109–120. Also, in Theory of probability and mathematical statistics, 66:119–132 (2003) (English version)
43. Roussas GG, Bhattacharya D (2007) Asymptotic expansions, exponential approximation and Hájek-Inagaki representation theorem under a general dependence set-up. In: Proceedings of the 20th Panhellenic statistical conference, Nikosia, Cyprus, April 11–15, 45–65 (Invited keynote lecture for the annual meeting)
44. Roussas GG, Bhattacharya D (2008) Hájek–Inagaki representation theorem, under general stochastic processes framework, based on stopping times. Stat Probab Lett 78:2503–2510
45. Roussas GG, Bhattacharya D (2009) Hájek–Inagaki convolution representation theorem for randomly stopped locally asymptotically mixed normal experiments. Stat Inference Stoch Process (In Print). DOI 10.1007/s 11203-008-9029-0. Published online: 16 December 2008
46. Roussas GG, Soms A (1973) On the exponential approximation of a family of probability measures and representation theorem of Hájek–Inagaki. Ann Inst Stat Math 25:27–39
47. Sweeting T (1992) Asymptotic ancillarity and conditional inference for stochastic processes. Ann Stat 20(1):580–589
48. Taniguchi M, Kakizawa Y (2000) Asymptotic theory of statistical inference for time series. Springer series in statistics. Springer, New York
49. van der Vaart AW (1998) Asymptotic statistics. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge
50. Wald A (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans Am Math Soc 54:426–482
51. Weiss L, Wolfowitz J (1966) Generalized maximum likelihood estimators. Theor Probab Appl 11:58–81
52. Weiss L, Wolfowitz J (1967) Maximum probability estimators. Ann Inst Stat Math 19:193–206
53. Weiss L, Wolfowitz J (1968) Generalized maximum likelihood estimators in a particular case. Theor Probab Appl 13:622–627
54. Weiss L, Wolfowitz J (1970) Maximum probability estimators and asymptotic sufficiency. Ann Inst Stat Math 22:225–244
55. Wolfowitz J (1965) Asymptotic efficiency of the maximum likelihood estimator. Theor Probab Appl 10:247–260

Chapter 18
Long Range Dependence in Third Order for Non-Gaussian Time Series

György Terdik

Abstract The object of this paper is to define the long-range dependence (LRD) for a Non-Gaussian time series in third order and to investigate the third order properties of some well known long-range dependent series. We define the third order LRD in terms of the third order cumulants and of the bispectrum. The definition of the third order LRD is given in polar coordinates.

18.1 Introduction

Higher order spectra and in particular bispectra are very efficient tools for time series analysis, the development of the rigorous theory has been started in the sixties [5,8], the applications include testing linearity and Gaussianity [14,31] parameter estimations [6, 19] and so on with a huge literature. At the same time many papers have been devoted to the research on the long memory time series. The memory of a time series is defined usually by the order of decay of the covariance function at infinity. The classical stationary ARMA models have short memory with exponentially decreasing covariance function. The long memory means that the decay is hyperbolic. The asymptotic distribution of the sums of the long memory processes is well studied due to [9, 10, 23–27, 32]. The dependence structure of a time series characterized by the covariances only for a Gaussian series, in this case being uncorrelated and independent is equivalent, hence long memory and long-range dependence are synonymous concepts in time series analysis. If the time series is not Gaussian all the higher order cumulants are necessary for the description of the dependence structure. The independence, for instance, implies and implied by that each higher order cumulant is zero, except all variables in the cumulant are the same. Although Non-Gaussianity and long-range dependence have been observed in many areas, in

G. Terdik
Department of Information Technology, Faculty of Informatics, University of Debrecen, Debrecen, Hungary
e-mail: TERDIK@delfin.UNIDEB.HU

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_18,

network traffic [22, 29, 37] in asset returns and exchange rate data [11], musics [7], etc., there are no results neither on the higher order statistics nor on higher order spectra of non-Gaussian and long-range dependent time series.

The object of this paper is defining the long-range dependence (LRD) of a time series in third order and investigating the third order properties of some well known long-range dependent non-Gaussian series. We define the third order LRD in terms of the third order cumulants and of the bispectrum. Both the third order cumulants and the bispectrum are symmetric, therefore we consider their values on the principal domains only. The definition of the third order LRD is given in polar coordinates, since the origin and the x-axis have some particular importance. Besides the primary singularity of the bispectrum at the origin, the singularity on the whole x-axis is allowed. Similarly, not only the radial decay of the third order cumulants at infinity is considered but its behavior when it is approaching the x-axis on the 'circle with infinite radius' as well. The marginal bispectrum and the third order cumulants on the x-axis couple these properties of the bispectrum and the third order cumulants. In Sect. 18.3 we consider five basic non-Gaussian models; the fractionally integrated noise, the linear fractional noise, H_2-process, Rosenblatt process and ΔLISDLG process serve as prototypes for the third order LRD. In the last section we summarize the results and put some conjectures for further subjects of investigations. The proofs are technical and omitted, interested readers may consult with [36] and the author.

18.2 Long-Range Dependence

It is a well known that for an i.i.d. series $X_1, X_2, \dots X_n$ (with mean zero and variance $\mathrm{Var}(X_k) = \sigma_X^2$) the variance of the sum of X_k's changes linearly with n, i.e.,

$$\mathrm{Var}\left(\sum_{k=1}^{n} X_k\right) = n\sigma_X^2.$$

Asymptotically the above result holds even when X_ℓ is a stationary time series. Let X_ℓ be a zero mean stationary time series given in spectral representation

$$X_\ell = \int_{-1/2}^{1/2} e^{i2\pi\ell\omega} \mathcal{M}_X(d\omega),$$

where $\mathsf{E}\left|\mathcal{M}_X(d\omega)\right|^2 = S_2(\omega)\, d\omega$, and $S_2(\omega)$ is its spectral density function. The variance of the sum $\sum_{\ell=0}^{n-1} X_\ell$ is calculated directly from the spectral representation, we have

$$\operatorname{Var}\left(\sum_{\ell=0}^{n-1} X_\ell\right) = \operatorname{Var}\left(\int_{-1/2}^{1/2} \frac{e^{i2\pi n\omega}-1}{e^{i2\pi\omega}-1}\mathcal{M}_X\left(d\omega\right)\right)$$
$$= \int_{-1/2}^{1/2} \left|\frac{e^{i2\pi n\omega}-1}{e^{i2\pi\omega}-1}\right|^2 S_2\left(\omega\right) d\omega.$$

Now, as far as $S_2\left(0\right)$ is finite the Fejér kernel $\left|\left(e^{i2\pi n\omega}-1\right)/\left(e^{i2\pi\omega}-1\right)\right|^2$ guarantees the limit

$$\lim_{n\to\infty}\frac{1}{n}\int_{-1/2}^{1/2} \left|\frac{e^{i2\pi n\omega}-1}{e^{i2\pi\omega}-1}\right|^2 S_2\left(\omega\right) d\omega = S_2\left(0\right),$$

hence for large n we obtain

$$\operatorname{Var}\left(\sum_{\ell=0}^{n-1} X_\ell\right) \simeq n S_2\left(0\right).$$

If this is true for a particular time series then it is expected that the estimated variances of the aggregated series for different n will show this approximation. Namely estimating and plotting the logarithm of these variances against $\log\left(n\right)$ we obtain a linear function with slope 1. Several phenomena have been reported in the literature, see [2,28], where the slope of $\log\left(n\right)$ is different from one, usually it is greater than 1, i.e., for large n

$$\log \operatorname{Var}\left(\sum_{k=1}^{n} X_k\right) \simeq 2H\log\left(n\right) + const.,$$

where $H \in (1/2, 1)$. Half-slope H is called the Hurst exponent, the name goes back to [17]. From now on we shall use a shifted version $h = H - 1/2$ of H, since it is frequently used in time series analysis. When the slope is larger than one, the spectrum has singularity at zero, which is equivalent to the divergence of the covariance series $\sum_{k=-\infty}^{\infty}\operatorname{Cov}\left(X_\ell, X_{\ell+k}\right)$. Therefore the decay of the covariances $\operatorname{Cov}\left(X_\ell, X_{\ell+k}\right)$ is slower, hence the dependence of the values X_ℓ and $X_{\ell+k}$ is stronger then in a usual case when the $S_2\left(0\right)$ is finite. A stationary time series X_ℓ, $\ell = 0, \pm1, \pm2 \ldots \pm n$ will be called *long-range dependent* (LRD) if its spectrum $S_2\left(\omega\right)$ behaves like $\left|\omega\right|^{-2h}$ at zero, more precisely

$$\lim_{\omega\to 0}\frac{S_2\left(\omega\right)}{\left|\omega\right|^{-2h} L\left(1/\left|\omega\right|\right)} = c_s, \tag{18.1}$$

where $h \in (0, 1/2)$ and $L\left(\cdot\right)$ is a slowly varying function at infinity. This definition of long-range dependence can be stated in terms of autocorrelation function as well, since, as far as $L\left(\cdot\right)$ is quasi-monoton slowly varying, (18.1) is equivalent to:

$$\lim_{k\to\infty} \frac{\mathrm{Cov}\,(X_{\ell+k}, X_\ell)}{|k|^{2h-1}\, L\,(k)} = \mathfrak{q}_c, \quad h \in (0, 1/2)\,. \tag{18.2}$$

Note here the connection between these two constants $\mathfrak{q}_s$ and $\mathfrak{q}_c$, namely the ratio $\mathfrak{q}_s/\mathfrak{q}_c = 2^{1-2h}\pi^{-2h}\Gamma\,(2h)\cos(\pi h)$ is constant. From now on this ratio will be denoted by γ_{2h}. In other words, the autocorrelation function decays hyperbolically. In fact although the spectrum is in L_1 its Fourier coefficients $\mathrm{Cov}\,(X_0, X_k)$ are not in L_1 any more. The equivalence of (18.1) and (18.2) is studied in the theory of regular varying functions in details, we quote here the results of two Theorems, 4.3.2 and 4.10.1 of [3]. Let $L\,(\cdot)$ be quasi-monotone slowly varying function, then for $h \in (0, 1/2)$

$$\sum_{k=-\infty}^{\infty} L\,(k)\,|k|^{2h-1}\, e^{i2\pi\omega k} \simeq 2^{1-2h}\pi^{-2h}\Gamma\,(2h)\cos(\pi h)\,|\omega|^{-2h}\, L\,(1/|\omega|)\,, \tag{18.3}$$

when $\omega \to 0$. The left hand side converges to a function, $S_2\,(\omega)$ say, such that it fulfils (18.1). The Tauberian conversion of this Abelian result is also true, namely if the Fourier coefficients of the right hand side, $\mathrm{Cov}\,(X_0, X_k)$ say, decreasing then it fulfils (18.2). These results of calculus covers totally the problem of LRD for Gaussian time series. It should be emphasized that either we are given the spectrum around zero

$$S_2(\omega) \simeq c_0\gamma_{2h}\,|\omega|^{-2h}\, L\,(1/|\omega|)\,, \quad \omega \to 0,$$

where c_0 is some (positive) constant which might depend on h, or the covariances at large k

$$\mathrm{Cov}(X_{\ell+k}, X_\ell) \simeq c_0\,|k|^{2h-1}\, L\,(|k|)\,, \quad k \to \infty,$$

they defines each other including the exponents, constants and slowly varying functions.

If the time series is not Gaussian then the higher order cumulants and spectra provide some additional information on the structure of dependence.

18.2.1 Bispectrum and Cumulants

Let the process X_ℓ be centered and stationary in third order, then its third order cumulants are

$$\begin{aligned}\mathrm{Cum}\left(X_{\ell+k_1}, X_{\ell+k_2}, X_\ell\right) &= \mathsf{E}X_{\ell+k_1}X_{\ell+k_2}X_\ell \\ &= C_3\,(k_{1:2})\,, \qquad k_1, k_2 = 0, \pm1, \pm2, \ldots\end{aligned}$$

(notation: $k_{1:2} = (k_1, k_2)$). The third order cumulants are called bicovariance as well. An easy consequence of this definition is that the following properties are fulfilled

$$\begin{aligned} C_3(k_{1:2}) &= C_3(k_2, k_1) \\ &= C_3(-k_1, k_2 - k_1). \end{aligned}$$

These equations provide the symmetry of the third order cumulants, the plain is divided into six equivalent parts, each of them is sufficient for the determining of the third order cumulants on the whole plane. One of these parts, called principal domain for the third order cumulants, is $0 \le k_2 \le k_1$, i.e., the lower half of the right upper quarter of the lattice with integer coordinates in the plain.

The bispectrum S_3 is a complex valued function of two variables with Fourier coefficients $C_3(k_{1:2})$, i.e.,

$$C_3(k_{1:2}) = \iint_{-1/2}^{1/2} e^{i2\pi(\omega_1 k_1 + \omega_2 k_2)} S_3(\omega_{1:2})\, d\omega_{1:2}.$$

In case the series $C_3(k_1, k_2)$ is in L_1 then the bispectrum has the series expansion

$$S_3(\omega_{1:2}) = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} C_3(k_{1:2})\, e^{-i2\pi(\omega_1 k_1 + \omega_2 k_2)}. \tag{18.4}$$

While the spectrum is real and nonnegative the bispectrum is generally complex valued and since $C_3(k_{1:2})$ is real, we have $S_3(\omega_{1:2}) = \overline{S_3}(-\omega_{1:2})$, (notation: $\omega_{1:2} = (\omega_1, \omega_2)$). The bispectrum S_3 is periodic

$$S_3(\omega_{1:2}) = S_3(\omega_1 + m_1, \omega_2 + m_2), \quad m_1, m_2 = \pm 1, \pm 2, \ldots,$$

and it has the following symmetry

$$\begin{aligned} S_3(\omega_{1:2}) &= S_3(\omega_2, \omega_1) \\ &= S_3(\omega_1, \omega_3). \end{aligned}$$

where $\omega_3 \triangleq -\omega_1 - \omega_2$. These symmetries imply 12 equivalent domains for the bispectrum, see Fig. 18.1, the principal domain traditionally is the triangle $\triangle_1$ with vertices (0,0), (1/2,0), and (1/3,1/3) ($\triangle_1$ denoted by T_1 in Fig. 18.1) [15, 35]. The Fig. 18.1 shows the symmetry lines going through on the origin for the bispectrum of a continuous time process as well.

The cumulants on the x-axis will play some particular roll later on, they are the Fourier coefficients of the *marginal bispectrum* M_3, i.e.,

$$\operatorname{Cum}(X_k, X_0, X_0) = \int_{-1/2}^{1/2} e^{i2\pi\omega k} M_3(\omega)\, d\omega,$$

The marginal bispectrum is calculated directly from the bispectrum

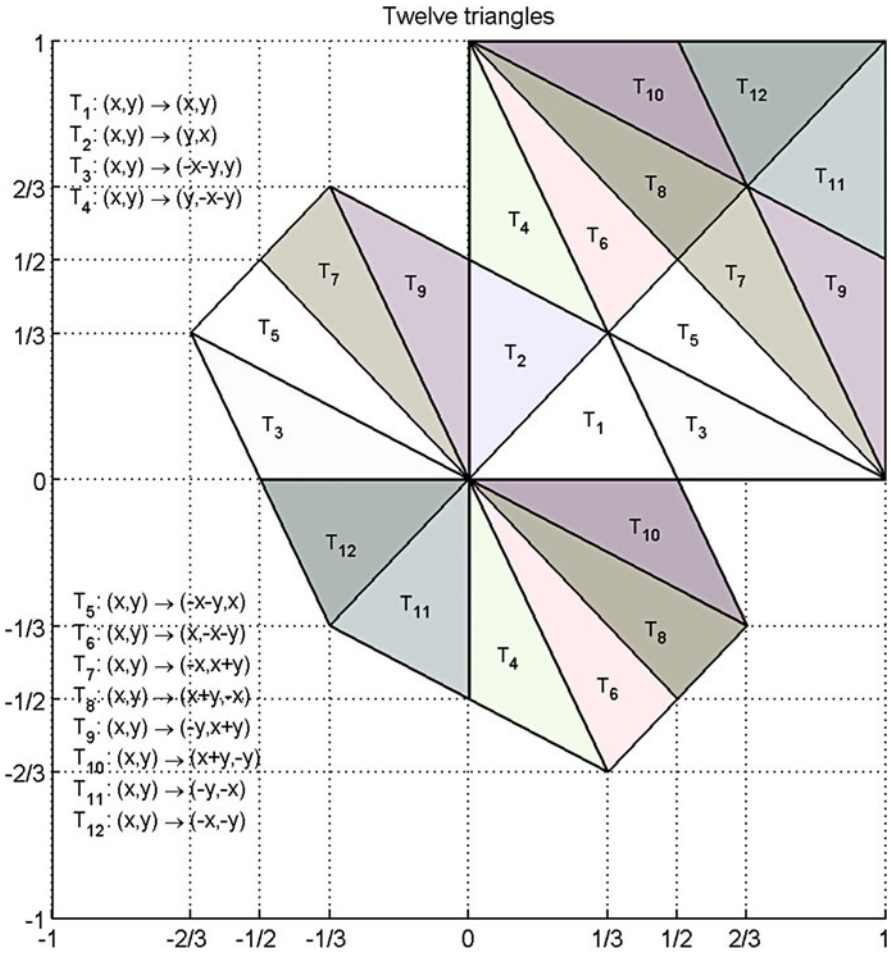


Fig. 18.1 12 domains of bispectrum

$$M_3(\omega) = \int_{-1/2}^{1/2} S_3(\omega, \lambda)\, d\lambda.$$

The following formula is useful for the calculation of the cumulant $C_3(\ell_{1:2})$ of a stationary process. When the bispectrum is given we have

$$C_3(\ell_{1:2}) = \int_{[0,1]^2} e^{i2\pi(\ell_1\omega_1+\ell_2\omega_2)} S_3(\omega_{1:2})\, d\omega_{1:2},$$

since the bispectrum is symmetric, see above we split the integral into 12 parts, each corresponds to a domain $\triangle_m, m = 1:12$, with boundary according to the symmetry lines, see Fig. 18.1. For instance $\triangle_1$ denotes the lower half of the upper quarter of

the plane, it corresponds to the triangle T_1 in the Fig. 18.1. Now, the transformation from $\triangle_m$ into $\triangle_1$ does not change the bispectrum (conjugates it at most) it changes the exponent only. Let us consider the integral

$$\int_{\triangle_5} e^{i\,2\pi(\ell_1\omega_1+\ell_2\omega_2)} S_3\left(\omega_{1:2}\right) d\omega_{1:2},$$

and transform $\triangle_5$ into $\triangle_1$, i.e., change the variables $\omega_1 = -\lambda_1 - \lambda_2$, and $\omega_2 = \lambda_1$, we obtain

$$\begin{aligned}\int_{\triangle_5} e^{i\,2\pi(\ell_1\omega_1+\ell_2\omega_2)} S_3\left(\omega_{1:2}\right) d\omega_{1:2} &= \int_{\triangle_1} e^{i\,2\pi([\ell_2-\ell_1]\lambda_1-\ell_1\lambda_2)} S_3\left(\lambda_{1:2}\right) d\lambda_{1:2} \\ &= c_3\left(\ell_2-\ell_1,-\ell_1\right),\end{aligned}$$

where

$$c_3\left(\ell_{1:2}\right) = \int_{\triangle_1} e^{i\,2\pi(\ell_1\omega_1+\ell_2\omega_2)} S_3\left(\omega_{1:2}\right) d\omega_{1:2}.$$

After transforming all domains into the $\triangle_1$ we obtain the following expression.

Lemma 18.1. *The continuous time cumulant function can be calculated in terms of integral c_3, namely*

$$C_3\left(\ell_1,\ell_2\right) = 4\,\mathrm{Re}\left(\operatorname*{sym}_{\ell_{1:2}}\left[c_3\left(\ell_1,\ell_2\right)+c_3\left(\ell_2-\ell_1,-\ell_1\right)+c_3\left(-\ell_1,\ell_2-\ell_1\right)\right]\right),$$

where $\operatorname{sym}_{\ell_{1:2}}$ *denotes the sum according to all possible permutations of* $\ell_{1:2}$ *divided by the number of the terms.*

18.2.2 Long-Range Dependence in Third Order

The relations (18.1) and (18.2) concern on the behavior of the covariances at infinity and the spectrum at zero. Similar results in 2D are available for isotropic case only, see [30, Ch. VII, Theorem 2.17].

Lemma 18.2. *Suppose that the bispectrum $S_3\left(\omega_{1:2}\right)$ is continuous. Then S_3 is isotropic if and only if $S_3\left(0,0\right)$ is finite and $S_3\left(\omega_{1:2}\right)$ is constant.*

Proof. Suppose that the bispectrum $S_3\left(\omega_{1:2}\right)$ is isotropic, i.e., $S_3\left(\omega_{1:2}\right)$ is constant on the circle with radius $\rho = |\omega_{1:2}|$, say. From the symmetry $S_3\left(\omega_{1:2}\right) = S_3\left(\omega_1,\omega_3\right)$ it follows $S_3\left(\rho,0\right) = S_3\left(\rho,-\rho\right)$. Hence the value of S_3 on the circle with radius ρ is the same as the value on the circle with radius $\sqrt{2}\rho$, equivalently $S_3\left(\rho\right) = S_3\left(\rho/\sqrt{2}\right) = \ldots = S_3\left(2^{-n/2}\rho\right)$, for any integer n. Now if $S_3\left(0\right)$ is finite then $S_3\left(\rho\right) = S_3\left(0\right)$ for any radius ρ. □

Since there is no isotropic bispectrum (except constant) we have to deal with more general 2D Fourier transforms than the isotropic one. Examples show below that the bispectrum $S_3(\omega_{1:2})$ and the third order cumulant are connected in some particular way. Let $\alpha_{\omega_2/\omega_1} = \arctan(\omega_2/\omega_1)$, i.e., α corresponds to the angle of the unit vector $\omega_{1:2}/|\omega_{1:2}|$ and the polar axis, if $\omega_{1:2}$ is in the principal domain when either the radius $\rho = |\omega_{1:2}| = \sqrt{\omega_1^2+\omega_2^2}$ tends to zero and $\alpha_{\omega_2/\omega_1}$ is fixed or $\alpha_{\omega_2/\omega_1}$ tends to zero and $|\omega_{1:2}|$ is fixed then the bispectrum might have singularity. When $\omega_{1:2}$ is close to the boundary we put the bispectrum in the following form

$$S_3(\omega_{1:2}) \simeq \mathfrak{p}_S |\omega_{1:2}|^{-3g_0} \alpha_{\omega_2/\omega_1}^{-2g_1} L_\alpha\left(|\omega_{1:2}|^{-1}\right) L\left(\alpha_{\omega_2/\omega_1}^{-1}\right), \tag{18.5}$$

where L is a slowly varying function and L_α is a slowly varying family. A term like $\left(\pi/4-\alpha_{\omega_2/\omega_1}\right)^{-2g}$ might be included into the bispectrum S_3, but we do not have any example implying its necessity.

Definition 18.1 (S). The time series X_ℓ is S-long-range dependent in third order with radial exponent g_0 and angular exponent g_1 if the bispectrum S_3 is factorized as (18.5) on the principal domain $\triangle_1$ and $0 < g_0 < 2/3$, $g_1 \in [0, 1/2)$.

If $g_1 > 0$ then the bispectrum has singularity on the hole x-axis therefore there are some additional information in the marginal bispectrum as well.

On the principal domain, $0 \le k_2 \le k_1$, let the third order cumulants be given in the form

$$\operatorname{Cum}\left(X_{k_1}, X_{k_2}, X_0\right) \simeq \mathfrak{p}_C |k_{1:2}|^{3g_0-2} L(|k_{1:2}|) K_\beta(|k_{1:2}|), \ |k_{1:2}| \to \infty \tag{18.6}$$

where $\beta_{k_2/k_1} = \arctan(k_2/k_1)$, $\beta_{k_2/k_1} \in (0, \pi/4)$, and for each $\beta \in (0, \pi/4)$, $K_\beta(\cdot)$ has a finite, continuous in β limit K_β, when $|k_{1:2}| \to \infty$, $L(|k_{1:2}|)$ is a slowly varying function. In addition,

$$K_\beta \simeq \mathfrak{p}_K \beta_{k_2/k_1}^{2g_2-1} \left(\pi/4-\beta_{k_2/k_1}\right)^{2g_3-1} L_1\left(\beta_{k_2/k_1}^{-1}\left(\pi/4-\beta_{k_2/k_1}\right)^{-1}\right), \tag{18.7}$$

when $\beta_{k_2/k_1} \to 0$. It is clearly corresponding to (18.5).

Definition 18.2 (C). The time series X_ℓ is C-long-range dependent in third order with radial exponent g_0 and angular exponent g_2, g_3 if the third order cumulants are factorized over the principal domain such that (18.6) and (18.7) fulfil, moreover $0 < g_0 < 2/3$, $g_2, g_3 \in (0, 1/2]$.

A delicate question to be addressed is the consequence of the LRD bispectrum to the marginal bispectrum

$$M_3(\omega) \simeq \mathfrak{c}\gamma_{2\tilde{h}} |\omega|^{-2\tilde{h}} L\left(|\omega|^{-1}\right), \tag{18.8}$$

around zero. The marginal cumulants, as k tends to infinity, are connected to M_3 in a well known way

$$\begin{aligned} \mathrm{Cum}\left(X_k, X_0, X_0\right) &\simeq \mathrm{c}\,|k|^{2\tilde{h}-1} L\left(|k|\right), \\ \mathrm{Cum}\left(X_{-k}, X_0, X_0\right) &\simeq |k|^{2\tilde{h}_{--}-1} L\left(|k|\right). \end{aligned} \tag{18.9}$$

Since $\mathrm{Cum}\left(X_{-k}, X_0, X_0\right) = \mathrm{Cum}\left(X_k, X_k, X_0\right)$, therefore this series corresponds to the cumulants on the diagonal $k_1 = k_2$.

18.3 Non-Gaussian LRD Models

There are several well known ways of modelling long-range dependence, see [1, 12, 34]. For instance one may start with either the fractional Gaussian noise process or with the discretized fractional Brownian motion. In this section we consider particular non-Gaussian processes, each of them are LRD, and derive the Hurst exponents for both the bispectra and cumulants.

18.3.1 Fractionally Integrated Noise

Define the fractional (not necessarily Gaussian) noise process as

$$\begin{aligned} U_\ell &= (1-B)^{-h} Z_\ell \\ &= \int_{-1/2}^{1/2} e^{i2\pi\omega\ell} \left(1 - e^{-i2\pi\omega}\right)^{-h} \mathcal{M}\left(d\omega\right), \end{aligned}$$

where Z_ℓ is i.i.d. series ($EZ_\ell = 0$, $EZ_\ell^2 = \sigma^2$) and B is the backward shift operator, i.e., $BZ_\ell = Z_{\ell-1}$. Note here that the stochastic spectral measure $\mathcal{M}\left(d\omega\right)$ has independent increments, $\mathsf{E}\left|\mathcal{M}\left(d\omega\right)\right|^2 = \sigma^2 d\omega$. Now, if $|h| < 1/2$, then U_ℓ is a stationary process having both moving average and infinite order AR representations, see [13, 16]. If $0 < h < 1/2$ then U_ℓ is a long-range dependent linear process in the sense that its autocorrelation function decreases hyperbolically (instead of the usual exponential rate). The moving average representation of U_ℓ is

$$U_\ell = \sum_{k=0}^{\infty} \frac{\Gamma\left(k+h\right)}{\Gamma\left(h\right)\Gamma\left(k+1\right)} Z_{\ell-k}. \tag{18.10}$$

The series (18.10) converges in mean-square. We have exact expressions for the covariances, they are is calculated by the MA representation (18.10), i.e.,

$$\mathrm{Cov}(U_{\ell+k}, U_\ell) = \sigma^2 \frac{\Gamma\left(k+h\right)\Gamma\left(1/2-h\right)}{2^{2h}\sqrt{\pi}\,\Gamma\left(k+1-h\right)\Gamma\left(h\right)}. \tag{18.11}$$

and $\sigma_U^2 = \sigma^2 \Gamma(1-2h)/\Gamma^2(1-h)$. For large values k we obtain the hyperbolic decay of correlations $(k \to \infty,)$

$$\begin{aligned} \mathrm{Cov}(U_{\ell+k}, U_\ell) &\simeq \sigma^2 \frac{\Gamma\,(1/2-h)}{2^{2h}\sqrt{\pi}\Gamma\,(h)} \,|k|^{2h-1} \\ &= \sigma^2 c_{2h}\,|k|^{2h-1}\,, \end{aligned} \tag{18.12}$$

where $c_{2h} = \Gamma\,(1-2h)\sin \pi h/\pi$. The spectral density follows from the spectral representation

$$S_{U,2}\,(\omega) = \sigma^2 \left|1 - e^{-i2\pi\omega}\right|^{-2h}. \tag{18.13}$$

The property (18.12) implies and is implied by the fact that the spectral density of U_ℓ is hyperbolic for low frequencies, i.e., if $\omega \to 0$, we obtain

$$\begin{aligned} S_{U,2}(\omega) &= \sigma^2\,|2\pi\omega|^{-2h} \left|\frac{e^{i2\pi\omega}-1}{2\pi\omega}\right|^{-2h} \\ &\simeq \sigma^2 \gamma_{2h} c_{2h}\,|\omega|^{-2h}\,. \end{aligned} \tag{18.14}$$

Vice versa, the corresponding covariance function writes as (18.12).

18.3.1.1 Third Order Properties

Assume Cum (Z_0, Z_0, Z_0) exists and denote it by $c_{Z,3}$, then

$$\mathrm{Cum}\,(U_0, U_0, U_0) = c_{Z,3} \frac{\Gamma\,(1+h/2)\,\Gamma\,(1-3h/2)}{\Gamma\,(1+h)\,\Gamma\,(1-h)\,\Gamma^2\,(1-h/2)}. \tag{18.15}$$

the third order cumulant Cum $\left(U_{k_1}, U_{k_2}, U_0\right)$ is expressed in terms of generalized Gauss function

$$\mathrm{Cum}\left(U_{k_1}, U_{k_2}, U_0\right) = \frac{c_{Z,3}}{\Gamma^2\,(h)} \frac{\Gamma\,(k_1+h)\,\Gamma\,(k_2+h)}{\Gamma\,(k_1+1)\,\Gamma\,(k_2+1)} A_{3,2}\,(h, k_{1:2})\,, \tag{18.16}$$

where

$$A_{3,2}\,(h, k_{1:2}) = {}_3F_2\,(h, k_1+h, k_2+h; k_1+1, k_2+1, 1)\,,$$

provided Re $(k_1+1+k_2+1-3h-k_1-k_2) > 0$, i.e., $h < 2/3$, in other words there are *no* extra assumptions for the third order analysis than the existence of the third order moment for Z_ℓ. We approximate $A_{3,2}\,(h, k_{1:2})$ by

$$r^h \frac{(\sin\beta)^h\,\Gamma\,(2-3h)}{\Gamma\,(2\,(1-h))}\;{}_2F_1\,(h, 1-h; 2\,(1-h)\,; 1-\tan\beta)\,.$$

If β is fixed and r tends to infinity then both k_1 and k_2 tend to infinity, hence

$$\operatorname{Cum}\left(U_{k_1}, U_{k_2}, U_0\right) \simeq c_{Z,3} r^{3h-2} K_\beta(r), \quad r \to \infty,$$

such that

$$\begin{aligned} K_\beta &\simeq (\sin\beta)^{2h-1} \frac{\Gamma(2-3h)}{\Gamma^2(h)\,\Gamma(2(1-h))} (\cos\beta)^{h-1} \\ &\quad \times {}_2F_1(h, 1-h; 2(1-h); 1-\tan\beta) \\ &= \beta^{2h-1} \frac{c_{2h}}{\Gamma(h)} L\left(\beta^{-1}\right). \end{aligned}$$

In other words

$$\lim_{\beta\to 0} \beta^{1-2h} \lim_{r\to\infty} \frac{\operatorname{Cum}\left(U_{k_1}, U_{k_2}, U_0\right)}{r^{3h-2}} = c_{Z,3} \frac{c_{2h}}{\Gamma(h)}. \tag{18.17}$$

The bispectrum $S_{U,3}$ is calculated in terms of the transfer function of a linear series

$$\begin{aligned} S_{U,3}(\omega_{1:2}) &= c_{Z,3} \prod_{j=1}^{3} \left(1 - e^{-i2\pi\omega_j}\right)^{-h} \\ &= (2\pi)^{-3h} c_{Z,3} \left[\Psi(\omega_{1:2})\right]^{-h} \prod_{k=1}^{3} (i\omega_k)^{-h}, \end{aligned} \tag{18.18}$$

here and the rest of the paper ω_3 is defined traditionally as $\omega_3 = -\omega_1 - \omega_2$, and

$$\Psi(\omega_{1:2}) = \prod_{j=1}^{3} \frac{\sin(\pi\omega_j)}{\pi\omega_j}.$$

As far as $\omega_1 \neq 0$ we have the representation for the bispectrum

$$S_{U,3}(\omega_{1:2}) \simeq |\omega_{1:2}|^{-3h} \alpha_{\omega_2/\omega_1}^{-h} (2\pi)^{-3h} e^{-ih\pi/2} L\left(\alpha_{\omega_2/\omega_1}^{-1}\right), \tag{18.19}$$

where $|\omega_{1:2}| = \rho$, $\alpha = \arctan(\omega_2/\omega_1)$.

18.3.1.2 Marginals

The cumulants behave differently on the boundaries

$$\operatorname{Cum}(U_k, U_0, U_0) \simeq c_{Z,3} \frac{\Gamma(1-2h)}{\Gamma(h)\,\Gamma^2(1-h)} k^{h-1}, \; k \to \infty, \tag{18.20}$$

$$\mathrm{Cum}\left(U_0, U_0, U_{-k}\right) \simeq c_{Z,3} \frac{\Gamma\left(2-3h\right)}{\Gamma\left(h\right)^2 \Gamma\left(2\left(1-h\right)\right)} k^{3h-2},\ k \to \infty.$$

Both series decay hyperbolically with different exponents.

The corresponding marginal bispectrum is

$$M_{U,3}\left(\omega\right) \simeq \frac{\Gamma\left(1-2h\right)}{\Gamma^2\left(1-h\right)} \left(2\pi\right)^{-h} e^{-i\pi h/2} \omega^{-h},\ \omega \to 0+,$$

since the Fourier transform according to the cumulants $\mathrm{Cum}\left(U_0, U_0, U_{-k}\right)$ on the diagonal provides singularity $|\omega|^{1-3h}$, while $\mathrm{Cum}\left(U_k, U_0, U_0\right)$ gives singularity $|\omega|^{-h}$, the superposition of these two sides results the singularity $|\omega|^{-h}$ of the marginal bispectrum, since $2-3h > 1-h$. When r is large the difference between the order of decays on the x-axis ($\mathrm{Cum}\left(U_k, U_0, U_0\right)$ with order $1-h$) and inside the principal domain ($\mathrm{Cum}\left(U_{k_1}, U_{k_2}, U_0\right)$ with order $2-3h$) is equalized, i.e., the radial exponent is $2-3h-(1-h) = 1-2h$.

Remark 18.1. For a linear process the bicoherence is constant,

$$\frac{\left|S_3\left(\omega_{1:2}\right)\right|^2}{S_2\left(\omega_1\right) S_2\left(\omega_2\right) S_2\left(\omega_3\right)} = const.,$$

hence from the singularity of the spectrum follows the singularity of the bispectrum.

18.3.2 Linear Fractional Noise

The continuous time spectrum

$$\left|\frac{e^{i2\pi\omega}-1}{i2\pi\omega}\right|^2 |2\pi\omega|^{-2h} = 4\sin^2\left(\pi\omega\right) |2\pi\omega|^{-2h-2}, \tag{18.21}$$

where $0 < h < 1/2$, behaves similarly, around zero, to the fractionally integrated spectrum (18.13). It corresponds to the continuous time process with spectral representation

$$V_t = \int_{-\infty}^{\infty} e^{i2\pi\omega t} \frac{e^{i2\pi\omega}-1}{i2\pi\omega} \left(i2\pi\omega\right)^{-h} \mathcal{M}\left(d\omega\right), \tag{18.22}$$

where $\mathcal{M}\left(d\omega\right)$ is a stochastic spectral measure with independent increments, $\mathsf{E}\left|\mathcal{M}\left(d\omega\right)\right|^2 = \sigma^2 d\omega$. We shall consider V_t at integer time points, $t = 0, \pm 1, \pm 2, \ldots$ only. If $\mathcal{M}\left(d\omega\right)$ is Gaussian then $Z_t = V_t$ is referred as (differenced) fractional Gaussian noise since $Z_t = \Delta B_t$, where B_t denotes the fractional Brownian motion [28]. We are interested in non-Gaussian case and call it fractional noise. The

spectrum of the fractional noise V_t follows from the Poisson summation formula, see [30],

$$S_{V,2}(\omega) = 4\sigma^2 (2\pi)^{-2h-2} \sin^2(\pi\omega) \sum_{m=-\infty}^{\infty} |m+\omega|^{-2h-2} \qquad (18.23)$$
$$= \sigma^2 |2\pi\omega|^{-2h} L\left(\omega^{-1}\right).$$

where

$$L\left(\omega^{-1}\right) = \left(\frac{\sin(\pi\omega)}{\pi\omega}\right)^2 \sum_{m=-\infty}^{\infty} \left|\frac{\omega}{m+\omega}\right|^{2h+2}.$$

The spectrum around zero

$$S_{V,2}(\omega) = (2\pi)^{-2h} \sigma^2 |\omega|^{-2h} L(1/|\omega|)$$
$$\simeq \sigma^2 c_{2h} \gamma_{2h} |\omega|^{-2h},$$

and the covariance

$$\mathrm{Cov}(V_{t+k}, V_t) = \sigma^2 \sigma_{2h} \Delta_{1/2}^2 k^{2h+1}$$
$$\simeq \sigma^2 c_{2h} |k|^{2h-1}.$$

where $\sigma_{2h} = c_{2h}/[2h(2h+1)]$.

18.3.2.1 Third Order Properties

The bispectrum

$$S_{V,3}(\omega_{1:2}) = c_{V,3} (2\pi)^{-3h} \Psi(\omega_{1:2}) \prod_{j=1}^{3} (i\omega_j)^{-h} \sum_{m_1,m_2=-\infty}^{\infty} \prod_{j=1}^{3} \left(\frac{\omega_j}{\omega_j + m_j}\right)^{h+1}. \qquad (18.24)$$

follows immediately from the transfer function of the process (18.22), assuming $\mathsf{E}\mathcal{M}(d\omega_1)\mathcal{M}(d\omega_2)\mathcal{M}(d\omega_3) = c_{V,3}\delta_{\omega_1+\omega_2+\omega_3} d\omega_{1:2}$. The bispectrum behaves around zero like

$$c_{V,3}\Psi(\omega_{1:2}) (i2\pi\omega_1)^{-h} (i2\pi\omega_2)^{-h} (i2\pi\omega_3)^{-h},$$

hence

$$S_{V,3}(\omega_{1:2}) \simeq c_{V,3}\rho^{-3h}\alpha^{-h} (2\pi)^{-3h} e^{-ih\pi/2} L\left(\alpha^{-1}\right),$$

where $\rho = |\omega_{1:2}|$, and $\alpha = \arctan(\omega_2/\omega_1)$. The third order properties, as far as $1/3 < h < 1/2$ and $\beta > 0$,

$$\operatorname{Cum}\left(V_{k_1}, V_{k_2}, V_0\right) \simeq c_{V,3} r^{3h-2} K_\beta(r) \beta^{2h-1} \frac{c_{2h}}{\Gamma(h)} L\left(\beta^{-1}\right), \; r \to \infty, \quad (18.25)$$

where $k_1 = r\cos\beta$ and $k_2 = r\sin\beta$, $\beta \in (0, \pi/4)$, moreover

$$K_\beta \simeq \beta^{2h-1} \frac{c_{2h}}{\Gamma(h)} L\left(\beta^{-1}\right), \quad \beta \to 0.$$

18.3.2.2 Marginals

Marginal cumulants are similar to those (18.20)

$$\operatorname{Cum}(V_k, V_0, V_0) \simeq 2k^{h-1}\sigma_{2h} \frac{c_{V,3}}{\Gamma(h)}, \quad (18.26)$$

$$\operatorname{Cum}(V_0, V_0, V_{-k}) \simeq k^{3h-2} \frac{\Gamma(2-3h)}{\Gamma(h)^2 \Gamma(2(1-h))}.$$

The marginal bispectrum is

$$M_{V,3}(\omega) \simeq 2c_{V,3}(2\pi)^{-h} \sigma_{2h} e^{-i\pi h/2} |\omega|^{-h}, \; \omega \to 0+.$$

18.3.3 H_2-Process

H_2-process is a homogenous Hermite process with order 2. The simplest process which is subject of non-central limit theorem. It is defined as the second order Hermite polynomial of the differenced fractional Brownian motion Z_t.

$$Z_t = \int_{-\infty}^{\infty} e^{i2\pi\omega t} \frac{e^{i2\pi\omega} - 1}{i2\pi\omega} (i2\pi\omega)^{-h} \mathcal{W}(d\omega).$$

The correlation of Z_t is the same as that of V_t, namely $R_Z(k) = \sigma_{2h}\Delta_{1/2}^2 k^{2h+1}$. Denote $H_2(Z) = Z^2 - \operatorname{Var}(Z)$, the Hermite polynomial with degree 2. The transform $X_t = H_2(Z_t)$ called H_2-process, has the following multiple Wiener–Ito representation

$$X_t = \int_{\mathbb{R}^2} e^{i2\pi t(\omega_1+\omega_2)} \frac{e^{i2\pi\omega_1} - 1}{i2\pi\omega_1} \frac{e^{i2\pi\omega_2} - 1}{i2\pi\omega_2} (i2\pi\omega_1)^{-h} (i2\pi\omega_2)^{-h} \mathcal{W}(d\omega_{1:2}).$$

The correlation of X_t is just the square of the correlation of Z_t (up to a constant), $\operatorname{Var}(Z_t) = \sigma_{2h}$,

$$R_X(k) = 2R_Z^2(k) = 2\sigma_{2h}^2\left(\Delta_{1/2}^2 k^{2h+1}\right)^2$$
$$\simeq 2c_{2h}^2 |k|^{2(2h-1)},$$

The spectrum around zero behaves like

$$S_{X,2}(\omega) = 2(2\pi)^{-4h}\int_{-\infty}^{\infty} |e_I(\lambda)|^2 |e_I(\omega-\lambda)|^2 |\lambda|^{-2h} |\omega-\lambda|^{-2h} d\lambda$$
$$\simeq 2c_{2h}^2\gamma_{4h-1}|\omega|^{1-4h}, \ \omega \to 0, \tag{18.27}$$

provided $h > 1/4$, where $e_I(\omega) = e^{i\pi\omega}\sin(\pi\omega)/\pi$.

The spectrum of $X_t = H_2(Z_t)$ is *LRD unless additionally* to $0 < h < 1/2$ we have $h > 1/4$. In general, it is easy to see that the $H_m(Z_t)$ will preserve the LRD property only if $h > (m-1)/2m$, since in this case the correlation $\rho_{H_m(Z)} = m!R_Z^m(k) \simeq m!\sigma_{2h}^m |k|^{m(2h-1)}$.

18.3.3.1 Third Order Properties

Concerning to the cumulants for Hermite polynomials, it follows from the general expression (see [35, p. 17])

$$\mathrm{Cum}\left(X_{k_1}, X_{k_2}, X_0\right) = 8R_Z(k_1)R_Z(k_1-k_2)R_Z(k_2) \tag{18.28}$$
$$= 8\sigma_{2h}^3\left[\Delta_{1/2}^2 k_1^{2h+1}\right]\left[\Delta_{1/2}^2(k_1-k_2)^{2h+1}\right]\left[\Delta_{1/2}^2 k_2^{2h+1}\right],$$

where $k_1 > k_2 > 0$, this cumulant is expressed also in terms of the bispectrum

$$\mathrm{Cum}\left(X_{k_1}, X_{k_2}, X_0\right) = \int_{\mathbb{R}^2} \exp\left(i2\pi k_1\omega_1 + k_2\omega_2\right) S_{X,3}(\omega_{1:2})\, d\omega_{1:2},$$

where

$$S_{X,3}(\omega_{1:2}) = 8(2\pi)^{-6h}\int_{-\infty}^{\infty} |e_I(\lambda)\, e_I(\omega_2-\lambda)\, e_I(\omega_1+\lambda)|^2 \tag{18.29}$$
$$\times |\lambda|^{-2h}|\omega_1-\lambda|^{-2h}|\omega_2+\lambda|^{-2h} d\lambda,$$

see [35, p. 47].

We are interested in the asymptotic behavior of the cumulants when k_1 and k_2 are large

$$\mathrm{Cum}\left(X_{t+k_1}, X_{t+k_2}, X_t\right) = 8R_Z(k_1)R_Z(k_1-k_2)R_Z(k_2)$$
$$\simeq 8c_{2h}^3\left(|k_1||k_1-k_2||k_2|\right)^{2h-1}.$$

For further notation the $\mathrm{Cum}\left(X_{t+k_1}, X_{t+k_2}, X_t\right)$ is denoted by $R_X(k_{1:2})$. First notice that if $k_1 = k_2$, or $k_2 = 0$, then $R_X(k_1, k_1) \simeq 8c_{2h}^2 |k_1|^{2(2h-1)}$. Put $k_1 = r\cos\beta$ and $k_2 = r\sin\beta$, $(\beta \neq \pi/4, 0)$, then

$$
\begin{aligned}
R_X(k_1, k_2) &\simeq 8c_{2h}^3 r^{3(2h-1)} K_\beta(r) \\
K_\beta &\simeq |\beta|^{2h-1} \left|\frac{\pi}{4} - \beta\right|^{2h-1} L(1/\beta) \\
L(1/\beta) &= |\cos\beta|^{2h-1} \left|\frac{\sin(\pi/4 - \beta)}{\pi/4 - \beta}\right|^{2h-1} \left|\frac{\sin\beta}{\beta}\right|^{2h-1},
\end{aligned}
$$

The singularity of the bispectrum (18.29) in accordance with the order of decay of R_X is the following. For small α we obtain

$$
S_{X,3}(\alpha, \rho) \simeq \rho^{1-6h} |\alpha|^{1-4h} L_\alpha(1/\rho) L(1/\alpha),
$$

it has radial singularity provided $h > 1/6$. The order of angular singularity of the bispectrum (α limit to 0) is $2h$ and cumulant has angular decay of order $1 - 2h$. The singularity of the bispectrum on the x-axis, according to the singularity of the angular part at 0, corresponds to the $2(2h-1)$ order of decay of R_X on the x-axis here the $S_3^{(s)}(\cdot)$ is the angular part, which is continuous on $(0, \pi/4)$, it has singularity at $\alpha = 0, \pi/4$ and by the symmetry at $\alpha = \pi/2, \pi, 5\pi/4$ on the unit circle, and $L(\cdot)$ is slowly varying at infinity. Cumulants

$$
R_X(k_{1:2}) \simeq \begin{cases} 8c_{2h}^3 r^{3(2h-1)} K_\beta(r), & k_1 \neq k_2, \\ 8c_{2h}^2 |k_1|^{2(2h-1)}, & k_2 = 0, k_1, \end{cases}, \quad r \to \infty.
$$

Again the order of decay of $R_X(k_{1:2})$ by the x-axis (and on the line $x = y$) is the difference $3(1-2h) - 2(1-2h) = 1 - 2h$ where $2(1-2h)$ is the radial singularity. The symmetry of the bispectrum implies that the values of the angular part on the unit circle are determined by the values on $[0, \pi/4]$.

The *divergence* of the 2D series of the third order cumulants is equivalent to the $2 - (3 - 6h) = 6h - 1$ order radial singularity of the bispectrum. The *assumption of the LRD* is $0 < 3 - 6h < 2$, i.e., $1/6 < h < 1/2$.

Lemma 18.3. *If $X_t = H_2(Z_t)$ and $h > 1/6$ then X_t is LRD in third order with Hurst exponents: $g_0 = (6h-1)/3$, $g_1 = 2h$, and $\tilde{h} = 4h - 1$.*

Corollary 18.1. *The assumption for the second and third order LRD is $1/4 < h < 1/2$. If $1/6 < h < 1/4$ then $X_t = H_2(Z_t)$ is LRD in third order but second order. If $0 < h < 1/6$ then the Gaussian time series Z_t is LRD but $X_t = H_2(Z_t)$ is not LRD neither in second nor in third order sense.*

18.3.3.2 Marginals

$$\begin{aligned}\mathrm{Cum}\left(X_k, X_0, X_0\right) &= 8R_Z^2\left(k\right)R_Z\left(0\right)\\ &= 8\sigma_{2h}^3\left[\Delta_{1/2}^2 k^{2h+1}\right]^2\\ &\simeq 8\sigma_{2h}c_{2h}^2\left|k\right|^{2(2h-1)}\end{aligned}$$

The marginal bispectrum

$$M_{X,3}\left(\omega_1\right) \simeq 8\sigma_{2h}c_{2h}^2\gamma_{4h-1}\left|\omega\right|^{1-4h},\ \omega \to 0. \tag{18.30}$$

18.3.4 Rosenblatt Process

Assume $1/4 < h < 1/2$, then

$$Y_t = \int_{\mathbb{R}^2} e^{i2\pi t(\omega_1+\omega_2)}\frac{e^{i2\pi(\omega_1+\omega_2)}-1}{i2\pi\left(\omega_1+\omega_2\right)}\left(i2\pi\omega_1\right)^{-h}\left(i2\pi\omega_2\right)^{-h}\mathcal{W}\left(d\omega_{1:2}\right), \tag{18.31}$$

defines a stationary process. The process Y_t, is subject of some non-central limit theorem [4, 9, 21, 33] it is called as Rosenblatt process. The representation (18.31) provides the covariance

$$\begin{aligned}\mathrm{Cov}\left(Y_{t+k}, Y_t\right) &= \frac{c_{2h}^2}{4h\left(4h-1\right)}\Delta_{1/2}^2 k^{4h} \qquad (18.32)\\ &\simeq 2c_{2h}^2\left|k\right|^{2(2h-1)},\quad k \to \infty,\end{aligned}$$

where $\Delta_{1/2}$ is the central difference operator and $\Delta_{1/2}^2$ is its square, in particular we have the variance

$$\mathrm{Var}\left(Y_t\right) = \frac{c_{2h}^2}{h\left(4h-1\right)}.$$

The spectrum of Y_t around zero is

$$S_{Y,2}\left(\omega\right) \simeq 2c_{2h}^2\gamma_{4h-1}\left|\omega\right|^{1-4h},\quad \omega \to 0, \tag{18.33}$$

provided $1/4 < h < 1/2$. We conclude that the asymptotic behavior of the spectra, $S_{X,2}$ and $S_{Y,2}$, at zero and the correlations at infinity are the same for both the series $H_2\left(Z_t\right)$ and its aggregated limit Y_t. This become quite obvious if we compare the spectra at zero and apply the well known result; the term $\left|\frac{e^{i2\pi\omega}-1}{i2\pi\omega}\right|^2$ has no singularity (≤ 1) in the interval $[-1/2, 1/2]$.

18.3.4.1 Third Order Properties

The bispectrum of the process Y_t follows from a general theory, see is

$$S_{Y,3}(\omega_{1:2}) = \Psi(\omega_{1:2})\, s_3(\omega_{1:2}),$$

where $\omega_3 = -\omega_1 - \omega_2$ as usual and where

$$s_3(\omega_{1:2}) = 8(2\pi)^{-6h} \int_{-\infty}^{\infty} |\lambda|^{-2h} |\omega_1 - \lambda|^{-2h} |\omega_2 + \lambda|^{-2h} d\lambda. \tag{18.34}$$

Note $h < 1/2$ implies $1 - 4h > -2h$, hence if $\omega_1 = \rho\cos\alpha$, $\omega_2 = \rho\sin\alpha$, we obtain

$$\int_{-\infty}^{\infty} |\lambda|^{-2h} |\omega_1 - \lambda|^{-2h} |\omega_2 + \lambda|^{-2h} d\lambda \simeq \rho^{1-6h} |\alpha|^{1-4h} L\left(\alpha^{-1}\right),$$

therefore

$$S_{Y,3}(\omega_{1:2}) \simeq \rho^{1-6h} |\alpha|^{1-4h} L\left(\alpha^{-1}\right) L_\alpha\left(\rho^{-1}\right).$$

Straightforward calculation shows

$$\begin{aligned}&\mathrm{Cum}\left(Y_{k_1}, Y_{k_2}, Y_0\right) = 8c_{2h}^3 \\ &\times \int_0^1 \int_0^1 \int_0^1 |k_1 + u_1 - u_3|^{2h-1} |k_2 + u_2 - u_3|^{2h-1} |k_1 - k_2 + u_1 - u_2|^{2h-1} du_{(1:3)}.\end{aligned}$$

It provides the limit for the third order cumulants

$$\lim_{r\to\infty} \frac{\mathrm{Cum}\left(Y_{k_1}, Y_{k_2}, Y_0\right)}{r^{6h-3}} = 8\frac{2^{h-1/2}}{\Gamma^3(2h)} L\left(\beta^{-1}(\pi/4-\beta)^{-1}\right) \beta^{2h-1} (\pi/4-\beta)^{2h-1},$$

$0 < 3 - 6h < 2$, i.e., $1/6 < h < 1/2$, hence one gets for K_β the expression

$$K_\beta = 8\frac{2^{h-1/2}}{\Gamma^3(2h)} \beta^{2h-1} (\pi/4-\beta)^{2h-1} L\left(\beta^{-1}(\pi/4-\beta)^{-1}\right).$$

18.3.4.2 Marginals

The cumulants on the x-axis and the diagonal are the same, i.e., if either $k_1 = 0$ or $k_1 = k_2 = k$, then

$$\mathrm{Cum}\left(Y_k, Y_k, Y_0\right) \simeq k^{2(2h-1)} \frac{8c_{2h}^3}{h(2h+1)}.$$

it follows that the marginal bispectrum is a spectrum, more over

$$M_{Y,3}(\omega) \simeq |\omega|^{1-4h} \frac{8c_{2h}^3}{h(2h+1)}\gamma_{4h-1},\ \omega \to 0. \tag{18.35}$$

18.3.5 ΔLISDLG Process

ΔLISDLG (increments of the Limit of Superpositions of Diffusion with Linear Generator) process, X_ℓ, $\ell \in \mathbb{Z}$, has been developed, see [18], in particular for modeling high speed network data. This discrete time process X_ℓ has several interesting properties, it is not self-similar rather multifractal, moreover it is nonlinear and LRD. Besides the Hurst exponent h, its distribution depends on two positive parameters c_0 and σ_0^2. Although all statistics of higher order has been derived in [18], we concern here the second and third order cumulants and spectra and repeat some results of [18].

- The spectral density of ΔLISDLG process X_ℓ is

$$\begin{aligned} S_{2,X}(\omega) &= \frac{\gamma_{2h}c_0\sigma_0^2}{1-2h}|\omega|^{-2h}\left|\frac{\sin\omega\pi}{\pi\omega}\right|^2 \sum_{m=-\infty}^{\infty}\left|\frac{\omega}{\omega+m}\right|^{2+2h} \\ &\simeq \frac{\gamma_{2h}c_0\sigma_0^2}{1-2h}|\omega|^{-2h},\ |\omega| \to 0. \end{aligned}$$

  for $\omega \in (0,1)$. The corresponding covariance function

$$\mathrm{Cov}(X_t, X_0) \simeq \frac{c_0\sigma_0^2}{1-2h}|t|^{2h-1},\ t \to \infty \tag{18.36}$$

- The bispectrum of X_t exists

$$S_{3,X}(\omega_{1:2}) = \frac{-3i(2\pi)^{-3-2h}c_0\sigma_0^4\prod_{j=1}^{3}\left(1-e^{i2\pi\omega_j}\right)}{\sin(\pi(h+1/2))\Gamma(2(1-h))}\sum_{m_{1:2}=-\infty}^{\infty}u(\omega_{1:2}+m_{1:2}),$$

  where

$$u(\omega_{1:2}) = \frac{\sum_{j=1}^{3}\omega_j|\omega_j|^{1-2h}}{(\omega_1\omega_2\omega_3)^2}.$$

  We are interested in the singularity of $S_{3,X}$ around zero. $S_{3,X}$ corresponds to the continuous time bispectrum

$$S_3(\omega_{1:2}) = \frac{-3ic_0\sigma_0^4(2\pi)^{-3-2h}\prod_{j=1}^{3}\left(1-e^{i2\pi\omega_j}\right)}{\sin(\pi(h+1/2))\Gamma(2(1-h))}u(\omega_{1:2})$$
$$= \frac{3\pi^{-1/2-2h}c_0\sigma_0^4\Gamma(1/2+h)}{(1-2h)\Gamma(1-h)}s_3(\omega_{1:2})\Psi(\omega_{1:2})$$

by Poisson summation formula, where

$$s_3(\omega_{1:2}) = \frac{\sum_{j=1}^{3}\omega_j|\omega_j|^{1-2h}}{\omega_1\omega_2\omega_3}$$
$$= \frac{\omega_1\left(|\omega_1|^{1-2h}-|\omega_3|^{1-2h}\right)+\omega_2\left(|\omega_2|^{1-2h}-|\omega_3|^{1-2h}\right)}{\omega_1\omega_2\omega_3}.$$

Hence the bispectrum is real-valued and positive, and around zero

$$s_3(\omega_{1:2}) \simeq 2(1-h)\rho^{-1-2h}L(1/\alpha),$$

more precisely

$$S_{3,X}(\omega_1,\omega_2) \underset{\rho\to 0}{\simeq} \frac{6\pi^{-1/2-2h}c_0\sigma_0^4(1-h)\Gamma(1/2+h)}{(1-2h)\Gamma(1-h)}\rho^{-1-2h}L(1/\alpha).$$

The interesting property is that $S_{3,X}$ has no singularity on the x-axis, nevertheless the bispectrum is LRD with exponent $g_0 = (1+2h)/3$ and $g_1 = 0$.

- Now, we turn to the third order cumulants,

$$\mathrm{Cum}\left(X_{t_1}, X_{t_2}, X_0\right) = c_3\,\Delta_{1/2}^2|t|^{1+2h}\Big|_{t=\mathrm{diam}(0,t_1,t_2)}, \tag{18.37}$$

for $t_1 \neq 0$, $t_2 \neq 0$, $t_1 \neq t_2$, respectively and $\mathrm{diam}(a,b,c) \triangleq \max(a,b,c) - \min(a,b,c)$ is the diameter, the constant

$$\mathrm{Cum}(X_0, X_0, X_0) = \frac{6c_0\sigma_0^4}{(h+1)(1-4h^2)},$$

is positive. If $t_1 \geq t_2 \geq 0$, then $\mathrm{diam}(t_1,t_2,0) = t_1$, hence

$$\mathrm{Cum}\left(X_{t_1}, X_{t_2}, X_0\right) \simeq \frac{6hc_0\sigma_0^4}{(1-2h)(h+1)}r^{2h-1}(\cos\beta)^{2h-1},$$

for large r, it corresponds to the singularity of S_3, namely $1-2h-2 = -1-2h$, [30, Ch. VII, Theorem 2.17].

- X_t is nonlinear.

- The discrete increment process X_t, $t \in \mathbb{Z}$, is not selfsimilar rather *multifractal*. Indeed the aggregated increments $X_t^{(n)}$, have cumulants of order m

$$\mathrm{Cum}_m\left(X_t^{(n)}\right) \simeq k_1(m) n^{2h-1}. \tag{18.38}$$

The Hurst exponent of n is $2h-1$, it does not depend on m. Assuming self-similarity, one obtains the second order cumulants (compare with (18.36))

$$\mathrm{Cum}_2\left(X_t^{(n)}\right) \simeq n^{2(h-1/2)}\,\mathrm{Cum}_2\left(X_t\right),$$

since the exponent follows from (18.36). Now, cumulants of order m follows from the selfsimilarity

$$\mathrm{Cum}_m\left(X_t^{(n)}\right) \simeq n^{m(h-1/2)}\,\mathrm{Cum}_m\left(X_t\right).$$

It implies the ΔLISDLG process X_t can not be self-similar rather h is changing with m in the above exponent of n, more precisely $h = h\,(m)$, such that

$$\mathrm{Cum}_m\left(X_t^{(n)}\right) \simeq n^{m(h(m)-1/2)}\,\mathrm{Cum}_m\left(X_t\right). \tag{18.39}$$

Let us compare the exponents in (18.38) and (18.39), we obtain

$$h(m) = \frac{2h-1}{m} + \frac{1}{2}.$$

If $m = 3$ then $h(3) = 2h/3 + 1/6$.

18.3.5.1 Marginals

The marginal cumulants have special form

$$\begin{aligned}\mathrm{Cum}\left(X_k, X_0, X_0\right) &= \mathrm{Cum}\left(X_k, X_k, X_0\right)\\ &\simeq \frac{6hc_0\sigma_0^4}{(1-2h)(h+1)}k^{2h-1},\end{aligned}$$

since $\mathrm{Cum}\left(X_k, X_0, X_0\right) = \mathrm{Cum}\left(X_0, X_0, X_{-k}\right)$, the marginal bispectrum is real and

$$M_{X,3}(\omega) \simeq \gamma_{2h}\frac{6hc_0\sigma_0^4}{(1-2h)(h+1)}|\omega|^{-2h},\ \omega \to 0.$$

Observe that the order of decays of cumulants inside the principal domain and on the boundaries are equal.

18.4 Conclusions

Although there are deep results on the Fourier transform in higher dimensions, see [20, 30, 38], we have not found some appropriate theory for problems arising in the theory of third order LRD time series. A question to be addressed is in what extent (Abelian and Tauberian type Theorems) the bispectrum (18.5) and cumulants (18.6) are connected to each other. Are Definition S and Definition C of third order LRD equivalent? From the above investigations of particular time series we make some conclusions.

Let us consider the results for the cumulants according to particular processes:

	Cumulants, decay				
	Marg: $\pi/4$	Ang: $\pi/4$	Radial	Ang: 0	Marg: 0
FrIN	0	$2-3h$	$2-3h$	$1-2h$	$1-h$
LinFrN	0	$2-3h$	$2-3h$	$1-2h$	$1-h$
H_2	$2\,(1-2h)$	$1-2h$	$3\,(1-2h)$	$1-2h$	$2\,(1-2h)$
Rosenblatt	$2\,(1-2h)$	$1-2h$	$3\,(1-2h)$	$1-2h$	$2\,(1-2h)$
ΔLISDLG	$1-2h$	0	$1-2h$	0	$1-2h$

This table implies that the sum of the angular exponent and the marginal exponent equals to the radial one. It follows that the exponents inside the principal domain is constant but they are changing continuously approaching to the x-axis on the circle with 'infinite radius'.

Conjecture 18.1. Let us consider the cumulants $\mathrm{Cum}\left(X_{k_1}, X_{k_2}, X_0\right)$ on the circle with 'infinite radius' and assume it is changing continuously, then approaching to both lines the x-axis ($\beta = 0$) and the diagonal ($\beta = \pi/4$) the orders of radial decays inside the triangle $(2-3g_0)$ and on the lines should be equalized by the angular exponents. For instance if $\beta = 0$,

$$(2-3g_0)-(1-2g_2) = 1-2\tilde{h},$$

therefore we have

$$\tilde{h} = \frac{3g_0-2g_2}{2}.$$

Now we turn to the bispectrum and consider the singularities

	Bispectrum, singularity				
	Marg: $\pi/4$	Ang: $\pi/4$	Radial	Ang: 0	Marg: 0
FrIN	$3h-1$	0	$3h$	h	h
LinFrN	$3h-1$	0	$3h$	h	h
H_2	$4h-1$	0	$6h-1$	$4h-1$	$4h-1$
Rosenblatt	$4h-1$	0	$6h-1$	$4h-1$	$4h-1$
ΔLISDLG	$2h$	0	$1+2h$	0	$2h$.

Conjecture 18.2 (Abel type). Let us suppose that the radial singularity of the bispectrum and the singularities of marginal bispectra are given. Then the radial decay of the cumulants and their angular 'decays' in both directions follow.

Conjecture 18.3 (Tauber type). Let us suppose that the radial decay of the cumulants and their angular 'decays' in both directions are given. Then the radial singularity of the bispectrum and the singularities of marginal bispectra follow.

References

1. Amblard PO, Brossier JM, Lacoume JL (1997) Playing with long range dependence and HOS. Proceedings of the IEEE signal processing, workshop on higher-order statistics, 1997, IEEE Xplore, pp 453–457
2. Beran J (1994) Statistics for long-memory processes. Monographs on statistics and applied probability, vol 61. Chapman and Hall, London
3. Bingham NH, Goldie CM, Teugels JL (1987) Regular variation. Encyclopedia of mathematics and its applications, vol 27. Cambridge University Press, Cambridge
4. Breuer P, Major P (1983) Central limit theorems for nonlinear functionals of Gaussian fields. J Multivar Anal 13(3):425–441
5. Brillinger DR (1965) An introduction to polyspectra. Ann Math Stat 36:1351–1374
6. Brillinger DR (1985) Fourier inference: Some methods for the analysis of array and nongaussian series data. Wat Resour Bull 21:743–756
7. Brillinger DR, Irizarry RA (1998) An investigation of the second- and higher-order spectra of music. Signal Process 65(2):161–179
8. Brillinger DR, Rosenblatt M (1967) Asymptotic theory of k-th order spectra. In: Harris B (ed) Spectral analysis of time series. Wiley, New York, pp 153–188
9. Dobrushin RL, Major P (1979) Non-central limit theorems for non-linear functionals of Gaussian fields. Z Wahrsch verw Gebiete 50:27–52
10. Doukhan P (2003) Models, inequalities, and limit theorems for stationary sequences. Theory and applications of long-range dependence. Birkhäuser, Boston, pp 43–100
11. Giraitis L, Robinson PM (2003) Parametric estimation under long-range dependence. Theory and applications of long-range dependence. Birkhäuser, Boston, pp 229–249
12. Granger CWJ, Ding Z (1996) Varieties of long memory models. J Econom 73(1):61–77
13. Granger CWJ, Joyeux R (1980) An introduction to long memory time series models and fractional differencing. J Time Ser Anal 1:15–30
14. Hinich MJ (1982) Testing for Gaussianity and linearity of a stationary time series. J Time Ser Anal 3L169–176
15. Hinich MJ, Messer H (1995) On the principal domain of the discrete bispectrum of a stationary signal. IEEE Trans Signal Process 43(9):2130–2134
16. Hosking JRM (1981) Fractional differencing. Biometrika 68:165–167
17. Hurst HE (1951) Long term storage capacity of reservoirs. Trans Am Soc Civil Eng 116: 770–808
18. Iglói E, Terdik Gy (2003) Superposition of diffusions with linear generator and its multifractal limit process. ESAIM Probab Stat 7:23–88 (electronic)
19. Jammalamadaka SR, Subba Rao T, Terdik Gy (2007) On multivariate nonlinear regression models with stationary correlated errors. J Stat Plan Inference R S N 137:3793–3814
20. Konovalov SP (1979) Absolute convergence of multiple Fourier series. Mat Zametki 25(2):211–216, 317
21. Major P (1981) Multiple Wiener–Itô integrals. Lecture notes in mathematics, vol 849. Springer, New York

22. Molnár S, Terdik Gy (2001) A general fractal model of internet traffic. In: Proceedengs of the 26th annual IEEE conference on local computer networks (LCN), Tampa, Florida, USA (Los Alamitos, CA, USA), IEEE Computer Society, IEEE, November 2001, pp 492–499
23. Rosenblatt M (1961) Independence and dependence. In: Proceedings of 4th Berkeley symposium on Mathematical Statistics and Probability, vol II. University of California Press, California, pp 431–443
24. Rosenblatt M (1976) Fractional integrals of stationary processes and the central limit theorem. J Appl Probab 13(4):723–732
25. Rosenblatt M (1979) Some limit theorems for partial sums of quadratic forms in stationary Gaussian variables. Z Wahrsch Verw Gebiete 49(2):125–132
26. Rosenblatt M (1981) Limit theorems for Fourier transforms of functionals of Gaussian sequences. Z Wahrsch Verw Gebiete 55(2):123–132
27. Rosenblatt M (1987) Remarks on limit theorems for nonlinear functionals of Gaussian sequences. Probab Theory Relat Fields 75(1):1–10
28. Samorodnitsky G, Taqqu MS (1994) Stable non-Gaussian random processes. Stochastic modeling, Stochastic models with infinite variance. Chapman and Hall, New York
29. Scherrer A, Larrieu N, Owezarski P, Borgnat P, Abry P (2007) Non-gaussian and long memory statistical characterizations for internet traffic with anomalies. IEEE Trans Dependable Secur Comput 4(1):56–70
30. Stein EM, Weiss G (1971) Introduction to Fourier analysis on Euclidean spaces. Princeton mathematical series, vol 32. Princeton University Press, Princeton
31. Subba Rao T, Gabr MM (1984) An introduction to bispectral analysis and Bilinear Time series. Lecture notes in statistics, vol 24. Springer, New York
32. Taqqu MS (1975) Weak convergence to fractional Brownian motion and to the Rosenblatt process. Z Wahrsch verw Gebiete 31:287–302
33. Taqqu MS (1979) Convergence of integrated processes of arbitrary Hermite rank. Z Wahrsch Verw Gebiete 50:53–83
34. Taqqu MS (2003) Fractional Brownian motion and long-range dependence. Theory and applications of long-range dependence. Birkhäuser, Boston, pp 5–38
35. Terdik Gy (1999) Bilinear stochastic models and related problems of nonlinear time series analysis; a frequency domain approach. Lecture notes in statistics, vol 142. Springer, New York
36. Terdik Gy (2008) Long-range dependence in higher order for non-gaussian time series. Acta Sci Math (Szeged) 74:663–684
37. Terdik Gy, Gyires T (2009) Lévy flights and fractal modeling of internet traffic. IEEE/ACM Trans Netw 17(1):120–129
38. Wainger S (1965) Special trigonometric series in k-dimensions. Mem Am Math Soc 59:102

Chapter 19
Graphical Models for Clustered Binary and Continuous Responses

Martin T. Wells and Dabao Zhang

Abstract Graphical models for clustered data mixed with discrete and continuous responses are developed. Discrete responses are assumed to be regulated by some latent continuous variables and particular link functions are used to describe the regulatory mechanisms. Inferential procedures are constructed using the full-information maximum likelihood estimation and observed/empirical Fisher information matrices. Implementation is carried out by stochastic versions of the generalized EM algorithm. As an illustrative application, clustered data from a developmental toxicity study is re-investigated using the directed graphical model and the proposed algorithms. A new interesting directed association between two mixed outcomes reveals. The proposed methods also apply to cross-sectional data with discrete and continuous responses.

19.1 Introduction

Analysis of clustered data with multidimensional outcomes has been intensively studied for decades. Research on multiple continuous outcomes has generated many useful models, e.g., seemingly unrelated regression models [32], simultaneous equation models [19], structural equation models [3] and multiple linear mixed models [23, 29, 30]. However, clustered data mixed with both discrete and continuous outcomes, which are common in epidemiological and medical studies, pose challenging issues in data analysis. A graphical association model proposed by [22], i.e., conditional Gaussian (CG) regression model, for cross-sectional study and [15] further extended it as hierarchical interaction model. An algorithm was developed by [16] for these models. In the case of clustered data, different strategies for modelling and inferencing correlated binary and normal responses have been proposed (see

M.T. Wells (✉)
Department of Statistical Science, Cornell University, 1190 Comstock Hall, Ithaca, NY 14853, USA
e-mail: mtw1@cornell.edu

M.T. Wells and A. SenGupta (eds.), *Advances in Directional and Linear Statistics*,
DOI 10.1007/978-3-7908-2628-9_19,

[13, 14, 17, 18], and [26]). All of these models focus on undirected association between variables representing discrete and continuous outcomes even though identifying directed associations may be primarily solicited. We may read off directed association from the CG-regression model and hierarchical interaction model. However the directed association may not, even in the case of cross-sectional study, be established from continuous variables to discrete variables.

In this article, we propose a general framework to model directed association between variables representing discrete and continuous outcomes in clustered data. Treating discrete outcomes as observed hidden continuous variables and using a latent graphical model to structurally construct directed association between continuous variables, this framework is an extension of the class of models developed by [17]. We propose a full-information maximum likelihood estimation procedure by developing several stochastic versions of the EM algorithm for the problem at hand. It is appealing to pursue Bayesian analysis with available powerful computers and "easy" implementation. However, the "easy" implementation of a Bayesian analysis usually takes advantage of Markov chain Monte Carlo (MCMC) approaches and follows some convenient priors. Unfortunately, it is impossible to guarantee the stationarity of a sampled chain, although some properties of stationarity can be checked (see review by [5]), and convenient priors may not yield proper posterior distributions (see review by [2]). Furthermore, it is well known that finding a noninformative prior for covariance structure is quite complex (see [10], and references contained therein) and since the models under study depend critically on the covariance structure, we focus our statistical inference on likelihood approach.

The building block of our model is the reciprocal graphical model in the sense of [21]. Consider the simultaneous equation system $Y = Y\Gamma + XB + U$ with G endogenous and K exogenous variables. One can associate a graph $\mathcal{G}$ with the simultaneous equation system. Specifically, $\mathcal{G} = (V, F)$ where $V = \{1, 2, \cdots, G + K\}$ is the set of vertices representing the random variables $Y_1, \cdots, Y_G, X_1, \cdots, X_K$ in this order, while the edge set F consists of the elements:

1. $(\alpha, \beta) \in F$ iff $1 \leq \beta \leq G$ and one of the following conditions hold:
   (a) $1 \leq \alpha \leq G; \alpha \neq \beta$ and $(\beta, \alpha) \notin Z_\Gamma$
   (b) $G + 1 \leq \alpha \leq G + K$ and $(\beta, \alpha - K) \notin Z_B$
2. $\{\alpha, \beta\} \in F$ iff $G + 1 \leq \alpha, \beta \leq G + K, \alpha \neq \beta$ and $(\alpha - G, \beta - G) \notin Z_\Phi$

where Z_Γ, Z_B, Z_Φ denote the sets of indices of all of the structural zeros of Γ, B and $\mathrm{Cov}(X)$, respectively. Since the class of undirected path components of $\mathcal{G}$ is $\{\{1\}, \cdots, \{G\}\} \cup \{u(\alpha) | G + 1 \leq \alpha \leq G + K\}$, where $u(\alpha)$ is undirected path component of α, one can see that there are no directed edges between the vertices which belong to the same path component. Hence $\mathcal{G}$ is a reciprocal graph as defined in [21]. It is shown in [21, Theorem 5.2], that under a set of regularity conditions that the probability distributions of the simultaneous equations system are global $\mathcal{G}$-Markov. In the case of recursive models (that is, Γ is lower triangular) we have the so called acyclic graphical Markov model.

We consider discrete outcomes as partially observed information of some latent continuous variables and incorporate the above reciprocal graphical model as a hidden model. Therefore, the directed association between mixed outcomes can be specified by constructing the association between (latent) continuous variables. This setup lets us extend and fully take the advantages of simultaneous equation model, in which prolific results have made in theory and application.

We illustrate the application of the models and algorithms by analyzing a dataset from a developmental toxicity study of ethylene glycol in mice conducted through the National Toxicity Program [27]. In a clustered case-control experiment, on each fetus two main outcomes of interest were collected as fetal weight (continuous) and fetal malformation status (binary). Previous studies have focused on the joint effects of ethylene glycol dose on fetal weight and on the probability of malformation. We further explore the possible directed association from fetal weight to fetal malformation status.

Our paper proceeds as follows. In Sect. 19.2, we introduce the graphical models for mixed and clustered responses under consideration. A full-information likelihood approach is considered in Sect. 19.3 by developing some stochastic versions of the EM algorithm for the models with known link functions. Section 19.4 provides an application of the proposed models and algorithms to the dataset in the developmental toxicity study. We confirm the intuitive directed association that fetus weighting less are more prone to malformation. Finally, a simulation study of the proposed algorithms is given in Sect. 19.5.

19.2 The Model

There is a long tradition in applied statistics to use continuous latent variables to model discrete variables, e.g., probit for binary data. Furthermore, limited-value data are also modelled by using continuous variables either because of existing latent censoring mechanism or just for convenience. For instance, in econometrics [20] introduced Tobit model for censored continuous variables based on the selection mechanism. Other types of Tobit models discussed in [1] further explore more complicated mechanisms between dependent censored variables and binary variables. These models are also useful in a variety of epidemiologic settings.

For multivariate outcomes $Z_{ij} = (Z_{1ij}\ Z_{2ij}\ \cdots\ Z_{Gij})$, $j = 1, 2, \cdots, m_i$, in the ith cluster, the kth variable Z_{kij} is connected to a group of latent continuous variables Y_{ij} by a link function $h_k(\cdot)$, i.e., $Z_{kij} = h_k(Y_{ij})$. In the simplest case, each latent continuous variable regulates only one observed outcome, i.e., $Z_{kij} = h_k(Y_{kij})$, and the continuous observed outcome has an identity function correspondingly. More complicated link functions can be used to describe the joint regulatory mechanisms between some outcomes such as in Tobit variants.

The directed/undirected association, within ith cluster, between components of Y_{ij} can be described by a reciprocal graph associated to the following simultaneous equations,

$$Y_{ij} = Y_{ij}\Gamma_i + X_{ij}B_i + U_{ij}, \quad U_{ij} \overset{iid}{\sim} N(0, \Sigma), \tag{19.1}$$

where $1 \leq j \leq m_i$; X_{it} is a K-dimensional row vector of observed exogenous variables; Γ_i and B_i are $G \times G$ and $K \times G$ respectively; the diagonal elements of Γ_i are zeros under normalization rule; and U_{ij} is a G-dimensional row vector of unobserved disturbances.

The above model describes within-cluster associations and therefore the coefficients matrices Γ_i and B_i may include cluster-dependent components. Putting all these cluster-dependent components as a column vector Θ_i. The between-cluster variation of Θ_i can be further modeled as,

$$\Theta_i = \eta_i \Theta + \xi_i, \quad \xi_i \overset{iid}{\sim} N(0, \Sigma),$$

where η_i includes all cluster-related characteristic affecting Θ_i. We focus on the simplest case with η_i as identity matrix, i.e., Θ_i is normally distributed with mean Θ and variance-covariance matrix Σ, although the general hierarchical model can also be considered following the same idea.

For notational simplicity, we further collect all cluster-independent unknown components of Γ_i and B_i into a column vector Ψ. With the re-specification of parameters, the above model associated to a reciprocal graph can be re-expressed in a much simpler form,

$$\begin{aligned} &Z_{kij} = h_k(Y_{ij}), 1 \leq k \leq G; \\ &\begin{cases} Y_{ij} = \Psi^T W_{ij} + \Theta_i^T M_{ij} + U_{ij}, \;\; U_{ij} \overset{iid}{\sim} N(0, \Sigma), \\ \Theta_i \overset{iid}{\sim} N(\Theta, V), \end{cases} \end{aligned} \tag{19.2}$$

where Θ and Ψ are P_1 and P_2-dimensional column vectors, respectively; M_{ij} and W_{ij} are $P_1 \times G$ and $P_2 \times G$ matrices, respectively, which are both constructed from elements of X_{ij} and Y_{ij}. Particularly, there are only P_1 nonzero components in M_{ij} and P_2 nonzero components in W_{ij}.

This formulation essentially gives a latent multivariate quasi-linear mixed model representation, which, however, is different from the multivariate linear mixed model because exogenous and endogenous variables may co-exist in both M_{ij} and W_{ij}. Since both M_{ij} and W_{ij} may depend on the disturbance term U_{ij}, we need to be careful in using the above form of the model for our inference. Note that the model in (19.2) is actually more general than in (19.1) because it can be derived from any simultaneous equation model linear in coefficients Γ_i and B_i without requiring Y_{ij} and X_{ij} to be linear in the right hand of the equations as in (19.1).

19.3 Full-Information Likelihood Inference

Inference on linear simultaneous equation models is usually considered either by an instrumental variable (IV) approach, e.g., two-step least square (2SLS) and three-step least square (3SLS), or by a likelihood approach, e.g., limited-information

maximum likelihood estimation (LIMLE) and full-information likelihood estimation (FIMLE). For the model proposed in the previous section, IV-type estimators are difficult to implement and, on the contrary, likelihood approaches are better suited for our implementation as shown in the following. An additional advantage of likelihood approaches is their statistical efficiency over the IV approaches.

Assume that there are n clusters in total, and, without loss of generality, that $m_i \equiv m$ for simplicity of notation. The joint density function of $Y_i = (Y_{i1}^T\ Y_{i2}^T\ \cdots\ Y_{im}^T)^T$ and Θ_i is

$$
\begin{aligned}
&(2\pi)^{-(Gm+P_1)/2}|\Sigma|^{-m/2}|V|^{-1/2}|I-\Gamma_i|^m \times \exp\left\{-\frac{1}{2}\Theta_i^T\left(V^{-1}\right.\right.\\
&\left.+\sum_{j=1}^{m} M_{ij}\Sigma^{-1}M_{ij}^T\right)\Theta_i+\Theta_i^T\left[\sum_{j=1}^{m} M_{ij}\Sigma^{-1}(Y_{ij}-\Psi^T W_{ij})^T+V^{-1}\Theta\right]\\
&\left.-\frac{1}{2}\sum_{j=1}^{m}(Y_{ij}-\Psi^T W_{ij})\Sigma^{-1}(Y_{ij}-\Psi^T W_{ij})^T-\frac{1}{2}\Theta^T V^{-1}\Theta\right\}. \qquad (19.3)
\end{aligned}
$$

Due to the complicated structures of M_{ij}, W_{ij} and Γ_i, integrating out Y_i and Θ_i and finding a closed-form of the marginal likelihood function may be hopeless. However, observing all Y_i and Θ_i will make it easy to estimate all parameters. Hence, we can consider (Y_i, Θ_i), $1 \le i \le n$, to be the complete data with observed data $Z_i = (Z_{i1}^T\ Z_{i2}^T\ \cdots\ Z_{im}^T)^T$, $1 \le i \le n$. Given the complete data representation, it is natural to pursue a generalized EM algorithm [12] to estimate the parameters.

19.3.1 The Generalized EM Algorithm

Let $\hat{\Theta}^{(k)}$, $\hat{\Psi}^{(k)}$, $\hat{V}^{(k)}$, $\hat{\Sigma}^{(k)}$ be the estimated parameters from the kth iteration of a generalized EM algorithm. And, for any function $g(Y_i, \Theta_i; \Theta, \Psi, V, \Sigma)$, define the iteration

$$
\begin{aligned}
&\mathrm{E}^{(k+1)}[g(Y_i, \Theta_i; \Theta, \Psi, V, \Sigma)|Z_i]\\
&\quad = \mathrm{E}[g(Y_i, \Theta_i; \Theta, \Psi, V, \Sigma)|Z_i, \hat{\Theta}^{(k)}, \hat{\Psi}^{(k)}, \hat{V}^{(k)}, \hat{\Sigma}^{(k)}]. \qquad (19.4)
\end{aligned}
$$

The expected complete-data log-likelihood in the $(k+1)$-st iteration is then

$$
\begin{aligned}
&Q(\Theta, \Psi, V, \Sigma; \hat{\Theta}^{(k)}, \hat{\Psi}^{(k)}, \hat{V}^{(k)}, \hat{\Sigma}^{(k)})\\
&\quad = -\frac{n}{2}(Gm+P_1)\log(2\pi)+\frac{nm}{2}\log|\Sigma^{-1}|+\frac{n}{2}\log|V^{-1}|\\
&\qquad + m\sum_{i=1}^{n}\mathrm{E}^{(k+1)}[\log|I-\Gamma_i||Z_i]-\frac{1}{2}\sum_{i=1}^{n}\mathrm{E}^{(k+1)}[(\Theta_i-\Theta)^T
\end{aligned}
$$

$$\times V^{-1}(\Theta_i - \Theta)|Z_i] - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{E}^{(k+1)}[(Y_{ij} - \Theta_i^T M_{ij}$$
$$- \Psi^T W_{ij})\Sigma^{-1}(Y_{ij} - \Theta_i^T M_{ij} - \Psi^T W_{ij})^T|Z_i]. \tag{19.5}$$

A further generalized EM algorithm can be constructed by maximizing the above expected complete-data log-likelihood function or searching for $(\Psi^{(k)}, V^{(k)}, \Sigma^{(k)})$ to improve it sufficiently at each iteration. However, the conditional expectations in the form of (19.4) may not have a closed-form expression. Since the joint density function of (Y_i, Θ_i) is available as in (19.3), we instead consider Monte Carlo approaches to evaluate these conditional expectations.

Based on different implementation strategies to evaluate the conditional expectation (19.4) using Monte Carlo samples, different stochastic versions of the generalized EM algorithm can be constructed. The Monte Carlo EM algorithm (MCEM) introduced by [31] evaluates this conditional expectation as the sample mean from some random draws of missing variables. Alternatively [4, 8], and [7] design the Stochastic EM algorithm (SEM) to avoid computation of the conditional expectations and impute only one draw of missing variables, conditional on observed data at each iteration. At each iteration, both MCEM and SEM algorithms update the estimates by imputing a new set of missing values and all the previously simulated missing values are then discarded, which is computationally inefficient. Instead, Delyon et al. [11] suggests the Stochastic Approximation EM algorithm (SAEM) to gradually discount the previously simulated missing values with a "forgetting factor" which is inversely proportional to the iteration step size. All three stochastic versions of the generalized EM algorithm are constructed for the model (19.2) in Appendix.

In the afore mentioned algorithms both Y_i and Θ_i must be sampled from their conditional distributions conditional on observed data, so we will call these algorithms full stochastic versions of the generalized EM algorithm hereafter. However, when $I - \Gamma_i$ does not include any random component, the stochastic-version generalized EM algorithms can be constructed without sampling the random coefficient Θ_i because all involved $\mathrm{E}^{(k+1)}[g(Y_i, \Theta_i; \Theta, \Psi, V, \Sigma)|Y_i]$ can be analytically expressed. In this case we thereafter refer to the corresponding generalized EM algorithms as the marginal stochastic versions, which can be developed following the same way as the full stochastic versions shown in Appendix.

19.3.2 Standard Errors

For notational simplicity, let Φ be a column vector including all unknown parameters in Ψ, Σ, Θ and V. Let $\mathcal{L}_c$ be the complete-data log-likelihood function. With Monte Carlo calculation of the conditional expectations (19.4), the standard errors of the point estimates can be calculated by using the observed Fisher information matrix, which can be approximated, using the missing information principle of [24],

by

$$\sum_{i=1}^{n} \hat{\mathrm{E}}^{(k+1)} \left[\frac{\partial}{\partial \Phi} \mathcal{L}_c(Y_i, \Phi_i) | Z_i \right] \times \sum_{i=1}^{n} \hat{\mathrm{E}}^{(k+1)} \left[\frac{\partial}{\partial \Phi^T} \mathcal{L}_c(Y_i, \Phi_i) | Z_i \right]$$
$$- \sum_{i=1}^{n} \hat{\mathrm{E}}^{(k+1)} \left[\frac{\partial^2}{\partial \Phi \partial \Phi^T} \mathcal{L}_c(Y_i, \Phi_i) + \frac{\partial}{\partial \Phi} \mathcal{L}_c(\Phi) \frac{\partial}{\partial \Phi^T} \mathcal{L}_c(Y_i, \Phi_i) | Z_i \right].$$

The first three conditional expectations are evaluated in the Expectation Step of the stochastic-version generalized EM algorithms. But the last conditional expectation is not, hence, extra computational effort is required.

The second-order derivatives of $\mathcal{L}_c$ and their conditional expectations are evaluated in calculating observed Fisher information matrix using missing information principle. However, the standard errors can be more easily evaluated by calculating empirical Fisher information matrix, which only requires the evaluation of the conditional expectation of the gradient vector of $\mathcal{L}_c$. When the data (Z_i, X_i, Θ_i) are independently and identically distributed, the corresponding expected Fisher information matrix can be approximated by the empirical Fisher information matrix [25],

$$\mathrm{I}_n(\Phi) = \sum_{i=1}^{n} \mathrm{E} \left[\frac{\partial}{\partial \Phi} \mathcal{L}_c(Y_i, \Theta_i; \Phi) | Z_i, \Phi \right] \cdot \mathrm{E} \left[\frac{\partial}{\partial \Phi} \mathcal{L}_c(Y_i, \Theta_i; \Phi) | Z_i, \Phi \right]^T.$$

All the involved conditional expectations can be stochastically approximated in the Expectation Step of the stochastic versions of the generalized EM algorithm.

19.4 The Developmental Toxicity Study

We will use the methodology proposed in the previous sections to analyze a dataset from the developmental toxicity study of ethylene glycol in mice conducted through the National Toxicity Program [27]. The experiment assigned pregnant mice randomly to four groups, one group serving as control and the others at different levels of ethylene glycol during major organogenesis and measurements of each fetus in the uterus were taken. The two main outcomes of interest on each fetus were continuously measured fetal weight and binary fetal malformation status. To investigate the joint effects of ethylene glycol dose on fetal weight and fetal malformation, many methods have been developed by jointly modelling the data with discrete and continuous outcomes [6, 9, 13, 14, 17, 28].

Here we are interested in checking the intuitive observation that a fetus weighting less is more prone to malformation, i.e., besides the usual associations between exogenous variables with fetal weight and fetal malformation status, there is a possible directed association from fetal weight to fetal malformation status. Thus, motivated by this intuition, we consider a directed graphical model described as

follows,

$$Z_{1ij} = Y_{1ij}, \;\; Z_{2ij} = I[Y_{2ij} > 0],$$
$$\begin{cases} Y_{1ij} = \beta_{1i} + \beta_2 X_{1ij} + \beta_3 X_{2ij} + \epsilon_{1ij}, \\ Y_{2ij} = \gamma_{1i} + \gamma_2 Y_{1ij} + \gamma_3 X_{1ij} + \epsilon_{2ij}, \end{cases}$$

where Z_{1ij} and Z_{2ij} are the weight and the malformation status of jth mouse in the ith litter, respectively; X_{1ij} and X_{2ij} are the dose administered and its quadratic term for the jth mouse in the ith litter, respectively. The error terms are distributed as independent bivariate normal distribution with the variance of ϵ_{1ij} as σ_{11} and the variance of ϵ_{2ij} as one for identification. The random coefficients $(\beta_{1i}, \gamma_{1i})$ are used to account for the cluster (or litter) effects and assumed to be distributed as a multivariate normal distribution with mean vector (β_1, γ_1) and variance-covariance matrix $V = (v_{ij})_{2\times 2}$.

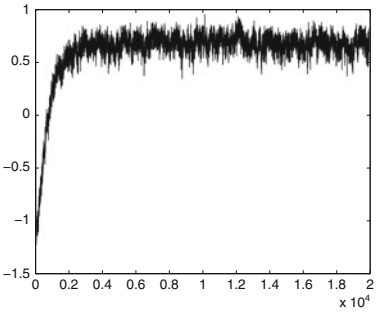
Starting from initial values with zero coefficients and identity variance-covariance matrices, we run 20,000 iterations of the marginal SEM algorithm to explore the possible values for the parameter estimation. The diagnostic convergence plots are of the estimated parameter series. Note that all chains are roughly stationary after 5,000 iterations (see Fig. 19.1). We then use the means of last 10,000 iterations in the estimated parameter chains as the initial values for our further inference.

With the available initial values, we run another 20,000 iterations of the marginal SAEM algorithm with $\gamma_k = 1/k$ but the first 50 iterations are implemented as the marginal MCEM algorithm with $l_k \equiv 50$. The diagnostic plots of the estimated parameter series demonstrated that 20,000 iterations are enough for us to get satisfactory point estimation for the parameters (see Fig. 19.1). The result of the marginal SAEM algorithm is summarized in Table 19.1, along with the corresponding standard errors (calculated by using the empirical Fisher information matrix).

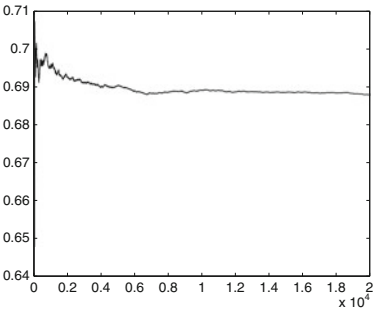
It is interesting to observe the complex association between fetal weight and fetal malformation. First, the result implies that the litter-specific effects on the fetal weight and malformation status are negatively correlated. Second, after accounting for the litter-specific effects and dose effects on both fetal weight and fetal malformation status, fetus weighting less are more prone to malformation. Therefore, in further investigation of ethylene glycol effects on fetal malformation it may better take into consideration the above sophisticated endogenous associations other than the simple "reduced" association.

19.5 Simulation Study

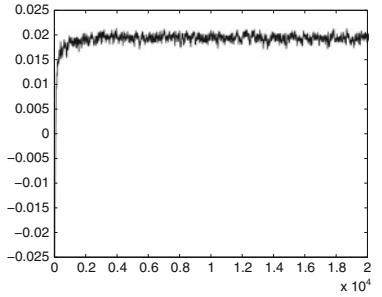
The performance of full and marginal stochastic versions of the generalized EM algorithm are studied by considering the same type of model as in the developmental toxicity study. Using a set of given parameters and the covariates in



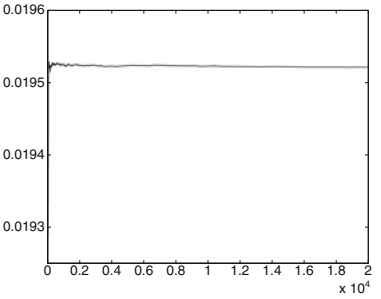
(a) SEM: γ_1



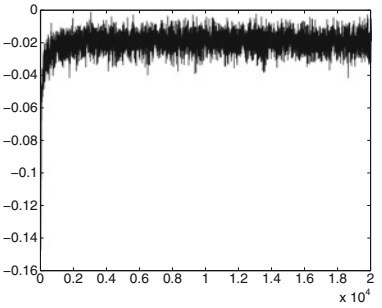
(a') SAEM: γ_1



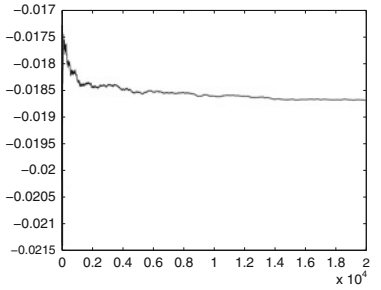
(b) SEM: β_3



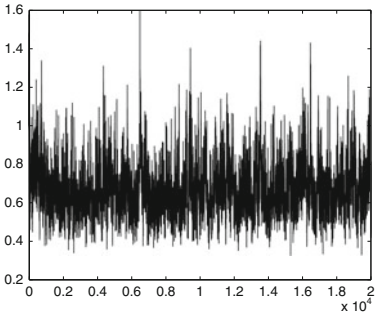
(b') SAEM: β_3



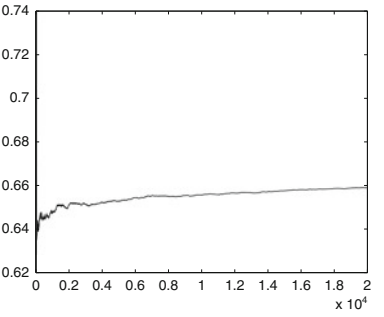
(c) SEM: v_{12}



(c') SAEM: v_{12}



(d) SEM: v_{22}



(d') SAEM: v_{22}

Fig. 19.1 Some convergence diagnostic plots for the marginal SEM and SAEM algorithms in the developmental toxicity study

Table 19.1 Parameter estimates in the developmental toxicity study

Parameter	Initial Value	Estimation Last 10,000	Estimation Lase one	S.E.	p-value Last One
$\hat{\beta}_1$	0.9723	0.9726	0.9726	0.0235	$\leq$0.0001
$\hat{\beta}_2$	−0.1470	−0.1476	−0.1476	0.3338	$\leq$ 0.0001
$\hat{\beta}_3$	0.0193	0.0195	0.0195	0.0098	0.0466
$\hat{\gamma}_1$	0.6794	0.6886	0.6880	0.3916	0.0789
$\hat{\gamma}_2$	−3.4707	−3.4573	−3.4575	0.3622	$\leq$ 0.0001
$\hat{\gamma}_3$	0.7615	0.7547	0.7548	0.0970	$\leq$ 0.0001
$\hat{\sigma}_{11}$	0.0056	0.0056	0.0056	0.0002	$\leq$0.0001
$\hat{\nu}_{11}$	0.0068	0.0068	0.0068	0.0012	$\leq$ 0.0001
$\hat{\nu}_{12}$	−0.0191	−0.0186	−0.0187	−0.0056	−0.0008
$\hat{\nu}_{22}$	0.6767	0.6576	0.6590	0.1909	0.0006

the developmental toxicity study, we simulate the observations for the endogenous variables. Since the proposed stochastic versions of the EM algorithm are computation-intensive, we perform the simulation study by using a fixed simulated data.

Following the same estimation strategy as in the developmental toxicity study, we run 20,000 iterations of the marginal SEM algorithm with zero coefficients and identity variance-covariance matrix as initial values (see Fig. 19.2). Then taking the means of the last 10,000 iterations in the estimated parameter series from the marginal SEM algorithm as the initial values. We also run another 20,000 iterations of marginal SAEM algorithm with $\gamma_k = 1/k$ but the first 50 iterations are implemented as the marginal MCEM algorithm with $l_k = 50$ (see Fig. 19.2). The results are shown in Table 19.2.

In addition, we also run 60,000 iterations of the full SEM algorithm with zero coefficients and identity variance-covariance matrix as initial values (see Fig. 19.3). Then taking the means of the last 30,000 iterations in the estimated parameter chains from the full SEM algorithm as the initial values, we run another 20,000 iterations of the full SAEM algorithm with $\gamma_k = 1/k$. However now the first 50 iterations are implemented as the full MCEM algorithm with $l_k = 50$ (see Fig. 19.3). The results are shown in Table 19.2. Comparing the upper panel and lower panel of Table 19.2, we observe that the estimation using full stochastic versions of the EM algorithms performs as well as that using marginal stochastic versions.

Comparing the plots in the left panel and those in the right panel of Fig. 19.3, we can also observe that some of the estimated parameter chains of the full SEM algorithm fluctuate in a much wider range than that of the marginal SEM algorithm. In particular, both estimated parameter chains for γ_1 and γ_2 of the full SEM algorithm have much higher autocorrelation than that of the marginal SEM algorithm. The rational for running more iterations in full SEM algorithm than that in marginal SEM algorithm analog is to explore the initial values for SAEM algorithms.

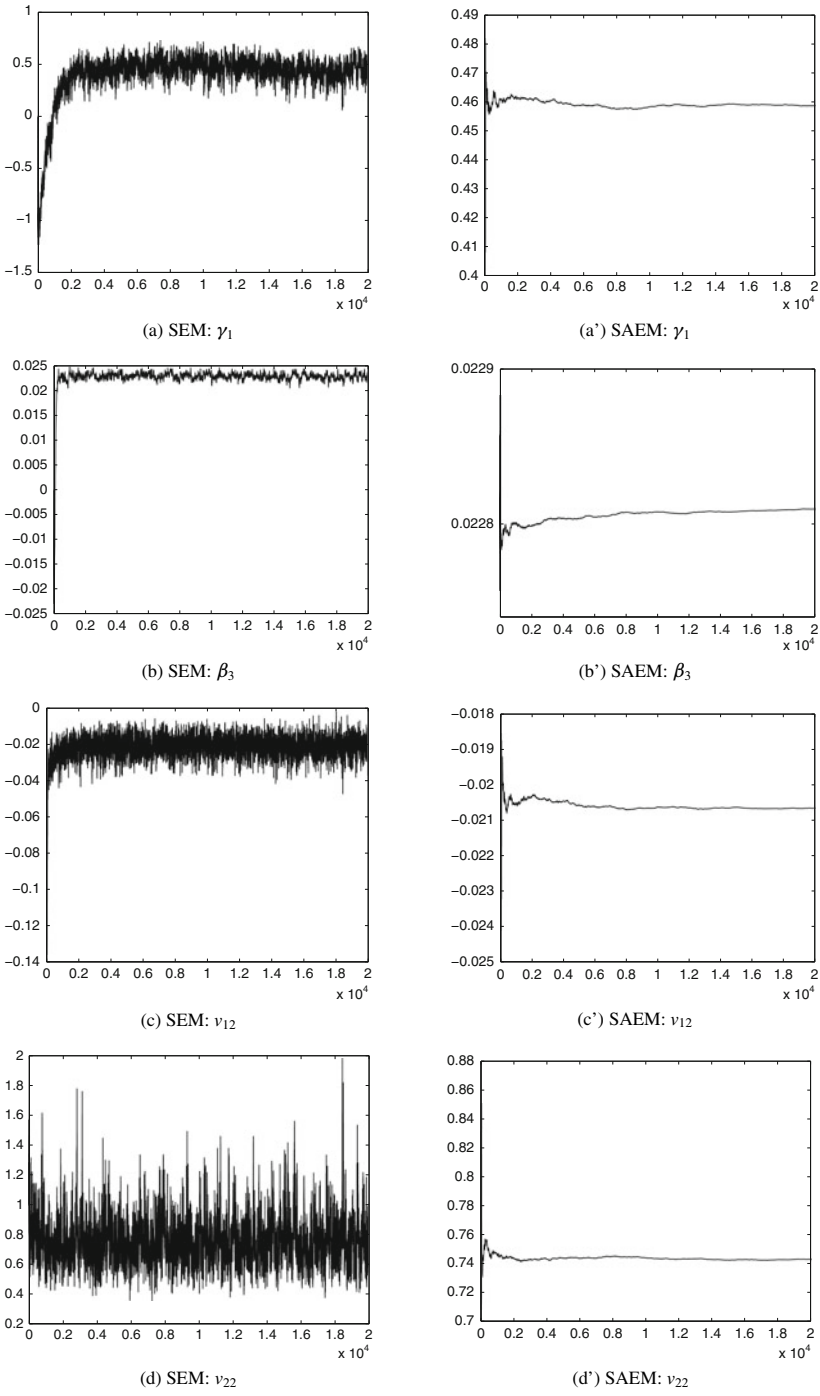


Fig. 19.2 Some convergence diagnostic plots for the marginal SEM and SAEM algorithms for the simulated data

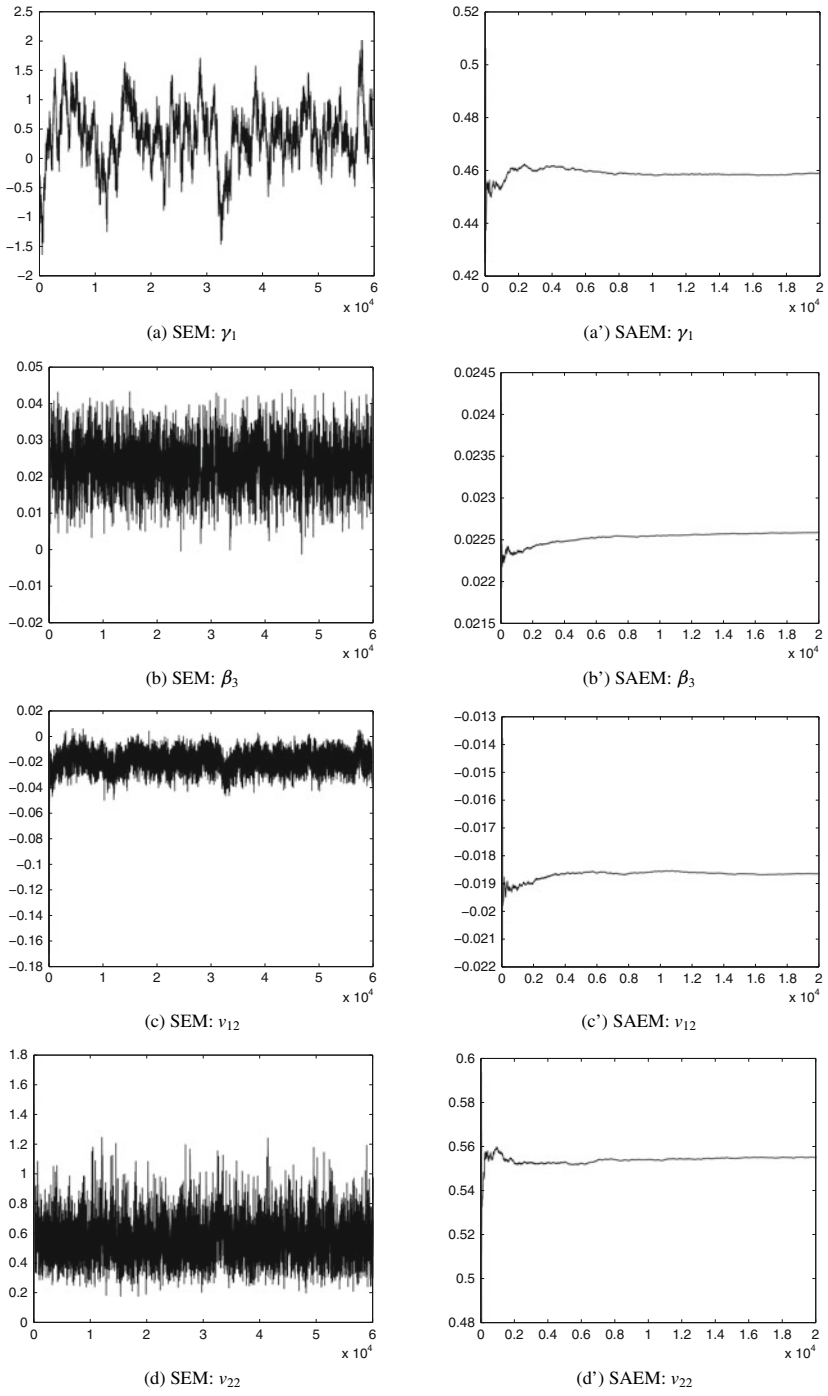
Table 19.2 Parameter estimates for the simulated data using marginal and full SAEM algorithms

	Parameter Value	True Value	Initial Last 10,000	Estimation Last One	Estimation	S.E.
Marginal	$\hat{\beta}_1$	0.9725	0.9884	0.9885	0.9885	0.0177
	$\hat{\beta}_2$	−0.1472	−0.1617	−0.1618	−0.1618	0.0278
	$\hat{\beta}_3$	0.0194	0.0228	0.0228	0.0228	0.0089
	$\hat{\gamma}_1$	0.6877	0.4504	0.4589	0.4588	0.2151
	$\hat{\gamma}_2$	−3.4585	−3.1904	−3.1852	−3.1852	0.3377
	$\hat{\gamma}_3$	0.7542	0.7989	0.7956	0.7956	0.1277
	$\hat{\sigma}_{11}$	0.0056	0.0054	0.0054	0.0054	0.0002
	$\hat{\nu}_{11}$	0.0068	0.0064	0.0064	0.0064	0.0011
	$\hat{\nu}_{12}$	−0.0190	−0.0211	−0.0207	−0.0207	0.0049
	$\hat{\nu}_{22}$	0.6657	0.7606	0.7433	0.7432	0.1844
Full	$\hat{\beta}_1$	0.9725	0.9892	0.9899	0.9899	0.0177
	$\hat{\beta}_2$	−0.1472	−0.1636	−0.1624	−0.1624	0.0276
	$\hat{\beta}_3$	0.0194	0.0234	0.0226	0.0226	0.0089
	$\hat{\gamma}_1$	0.6877	0.4480	0.4587	0.4591	01782
	$\hat{\gamma}_2$	−3.4585	−209465	2.9634	2.9634	0.3237
	$\hat{\gamma}_3$	0.7542	0.7289	0.7293	0.7293	0.1129
	$\hat{\sigma}_{11}$	0.0056	0.0054	0.0054	0.0054	0.0002
	$\hat{\nu}_{11}$	0.0068	0.0065	0.0064	0.0064	0.0011
	$\hat{\nu}_{12}$	−0.0190	−0.0189	−0.0186	−0.0186	0.0043
	$\hat{\nu}_{22}$	0.6657	0.5538	0.5547	0.5551	0.1417

19.6 Discussion

In this article we use a latent reciprocal graphical model to construct a directed graphical model for clustered data mixed with discrete and continuous variables. Discrete variables are assumed to be versions of some latent continuous variables and the discretization mechanisms are described by known link functions. In more general cases, such as ordered discrete variables and Box-Cox models, we may have to use link functions but with unknown parameters to bridge observed variables and latent variables used for constructing the hidden reciprocal graphical models, i.e., $Z_{kij} = h_k(Y_{ij}, \Xi)$. In the case with the ordinal discrete variables, the threshold parameters (included in the link parameter Ξ) may not be updated in each iteration of EM algorithms by maximizing the expected complete-data log-likelihood since they define the integration domains. So, some other strategies may have to be pursued. However, Box-Cox models can be put into a much more general model with known link function, for clustered data,

$$
\begin{aligned}
& Z_{kij} = h_k(Y_{ij}), 1 \le k \le G; \\
& \begin{cases} H(Y_{ij}, X_{ij}, \Theta_i, \Psi) = U_{ij}, \;\; U_{ij} \overset{iid}{\sim} N(0, \Sigma), \\ \Theta_i = \eta_i \Theta + \xi_i, \;\; \xi_i \overset{iid}{\sim} N(0, V), \end{cases}
\end{aligned}
$$



(a) SEM: γ_1 (a') SAEM: γ_1

(b) SEM: β_3 (b') SAEM: β_3

(c) SEM: v_{12} (c') SAEM: v_{12}

(d) SEM: v_{22} (d') SAEM: v_{22}

Fig. 19.3 Some convergence diagnostic plots for the full SEM and SAEM algorithms for the simulated data

where the random coefficients in Θ_i are further regressed on some panel-specific covariates η_i. Then the parameters in the Box-Cox transformation are included in Ψ and the corresponding link functions are set as identity function. The algorithms proposed here may be modified for a broad class of function $H(\cdot)$.

As shown in the simulation study and the developmental toxicity study, the SEM algorithm provides fluctuating estimates at each iteration but the SAEM algorithm usually gets trapped and steadily converges at some local points close to the initial values. Therefore, we propose to explore initial values by using the SEM algorithm and then run the SAEM algorithm to get point estimation of the parameters and to calculate the standard errors of the estimators. Although we have only focused clustered data in this article, all the models and algorithms can be easily modified for a cross-sectional study.

Acknowledgements The authors thank R.V. Gueorguieva and A. Agresti for the data use of the Developmental Toxicity Study. Wells was partially supported by NSF grant 06-12031 and NIH grant R01-GM083606-01.

Appendix

Stochastic Versions of the Generalized EM Algorithms

Stochastic versions of the generalized EM algorithm essentially evaluates the conditional expectations (19.4) by $\hat{\mathrm{E}}^{(k+1)}[g(Y_i, \Theta_i; \Theta, \Psi, V, \Sigma)|Z_i]$ using various Monte Carlo approaches. The expected complete-data log-likelihood function (19.5) can then be stochastically approximated by

$$\begin{aligned}
&\hat{Q}^{(k+1)}(\Theta, \Psi, V, \Sigma; \hat{\Theta}^{(k)}, \hat{\Psi}^{(k)}, \hat{V}^{(k)}, \hat{\Sigma}^{(k)}) \\
&= -\frac{n}{2}(Gm + P_1)\log(2\pi) + \frac{nm}{2}\log|\Sigma^{-1}| + \frac{n}{2}\log|V^{-1}| \\
&\quad + m\sum_{i=1}^{n}\hat{\mathrm{E}}^{(k+1)}\Bigg[\log|I - \Gamma_i||Z_i] - \frac{1}{2}\sum_{i=1}^{n}\hat{\mathrm{E}}^{(k+1)}[(\Theta_i - \Theta)^T \\
&\quad \times V^{-1}(\Theta_i - \Theta)|Z_i] - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\mathrm{E}}^{(k+1)}[(Y_{ij} - \Theta_i^T M_{ij} \\
&\quad - \Psi^T W_{ij})\Sigma^{-1}(Y_{ij} - \Theta_i^T M_{ij} - \Psi^T W_{ij})^T|Z_i\Bigg].
\end{aligned}$$

Maximizing $\hat{Q}^{(k+1)}(\Theta, \Psi, V, \Sigma; \hat{\Theta}^{(k)}, \hat{\Psi}^{(k)}, \hat{V}^{(k)}, \hat{\Sigma}^{(k)})$ for each of Θ,Ψ,V and Σ separately, we have closed-from estimators for Θ and V; but Ψ and Σ have no closed-form estimators. Therefore, we set $\Psi = \hat{\Psi}^{(k)}$ to maximize

$$\hat{Q}^{(k+1)}(\hat{\Theta}^{(k+1)}, \hat{\Psi}^{(k)}, \hat{V}^{(k+1)}, \Sigma; \hat{\Theta}^{(k)}, \hat{\Psi}^{(k)}, \hat{V}^{(k)}, \hat{\Sigma}^{(k)})$$

and get a closed-form estimator for Σ. Then we maximize $\hat{Q}^{(k+1)}(\hat{\Theta}^{(k+1)}, \Psi, \hat{V}^{(k+1)}, \hat{\Sigma}^{(k+1)}; \hat{\Theta}^{(k)}, \hat{\Psi}^{(k)}, \hat{V}^{(k)}, \hat{\Sigma}^{(k)})$ for $\hat{\Psi}^{(k+1)}$. The stochastic version of the generalized EM algorithm can therefore be constructed as follows.

Expectation Step For $1 \le i \le n$, calculate

$$\begin{cases} \hat{\mathrm{E}}^{(k+1)}[\Theta_i | Z_i], \\ \hat{\mathrm{E}}^{(k+1)}[\Theta_i \Theta_i^T | Z_i], \\ \hat{\mathrm{E}}^{(k+1)}[\log |I - \Gamma_i| | Z_i], \\ \sum_{j=1}^m \hat{\mathrm{E}}^{(k+1)}[\mathrm{vec}(W_{ij}) \otimes \mathrm{vec}(W_{ij}) | Z_i], \\ \sum_{j=1}^m \hat{\mathrm{E}}^{(k+1)}[(Y_{ij} - \Theta_i^T M_{ij})^T \otimes \mathrm{vec}(W_{ij}) | Z_i], \\ \sum_{j=1}^m \hat{\mathrm{E}}^{(k+1)}[(Y_{ij} - \Theta_i^T M_{ij}) \otimes (Y_{ij} - \Theta_i^T M_{ij}) | Z_i]; \end{cases}$$

Maximization Step Calculate

$$\begin{cases} \hat{\Theta}^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \hat{\mathrm{E}}^{(k+1)}[\Theta_i | Z_i], \\ \hat{V}^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \hat{\mathrm{E}}^{(k+1)}[\Theta_i \Theta_i^T | Z_i] - \hat{\Theta}^{(k+1)} \hat{\Theta}^{(k+1)T}, \\ \hat{\Sigma}^{(k+1)} = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \hat{\mathrm{E}}^{(k+1)}[(Y_{ij} - \Theta_i^T M_{ij} - \Psi^T W_{ij})^T \\ \quad \times (Y_{ij} - \Theta_i^T M_{ij} - \Psi^T W_{ij}) | Z_i]|_{\Psi = \hat{\Psi}^{(k)}}, \\ \hat{\Psi}^{(k+1)} = \mathrm{argmax}_\Psi \{ m \sum_{i=1}^n \hat{\mathrm{E}}^{(k+1)}[\log |I - \Gamma_i| | Z_i] \\ \quad - \frac{1}{2} \Psi^T \sum_{i=1}^n \sum_{j=1}^m \hat{\mathrm{E}}^{(k+1)}[W_{ij} \Sigma^{-1} W_{ij}^T | Z_i] \Psi \\ \quad + \sum_{i=1}^n \sum_{j=1}^m \hat{\mathrm{E}}^{(k+1)}[(Y_{ij} - \Theta_i^T M_{ij}) \Sigma^{-1} W_{ij}^T | Z_i] \Psi \}|_{\Sigma = \hat{\Sigma}^{(k+1)}}; \end{cases}$$

where $\otimes$ is the Kronecker product, and $\mathrm{vec}(W_{ij})$ collects all of the nonzero components of W_{ij} to form a $P_2 \times 1$ vector. All of the conditional expectations used in the Maximization Step are calculated in the Expectation Step. Note that in general the stochastic approximation of $\hat{\mathrm{E}}^{(k+1)}[\log |I - \Gamma_i| | Z_i]$ may involve unknown parameters. A solution of this is to consider a series approximation approach and expand $\log |I - \Gamma_i|$ first into summation and then evaluate its each summand. In the case that $|I - \Gamma_i|$ does not involve any random coefficient in Θ_i, this dependency is not an issue.

In the case $|I - \Gamma_1|$ involves Ψ, we have to consider nonlinear optimization routines for $\hat{\Psi}^{(k+1)}$. One-step Newton–Raphson algorithm may be used to update $\hat{\Psi}^{(k)}$ to $\hat{\Psi}^{(k+1)}$. If $|I - \Gamma_1|$ does not involve Ψ, then $\hat{\Psi}^{(k+1)}$ can be calculated explicitly as

$$\hat{\Psi}^{(k+1)} = \left(\sum_{i=1}^n \sum_{j=1}^m \hat{\mathrm{E}}^{(k+1)}[W_{ij} \Sigma^{-1} W_{ij}^T | Z_i] \right)^{-1} \times \sum_{i=1}^n \sum_{j=1}^m \hat{\mathrm{E}}^{(k+1)}[W_{ij} \Sigma^{-1} (Y_{ij} - \Theta_i^T M_{ij}) | Z_i]|_{\Sigma = \hat{\Sigma}^{(k+1)}}.$$

A.1 MCEM Algorithm

The conditional expectations (19.4) are evaluated by using l_{k+1} ($l_{k+1} \gg 1$) random draws of missing variables (Y_i, Θ_i) conditional on observable data Z_i, i.e.,

$$\hat{\mathrm{E}}^{(k+1)}[g(Y_i, \Theta_i; \Theta, \Psi, V, \Sigma)|Z_i] = \frac{1}{l_{k+1}} \sum_{t=1}^{l_{k+1}} g(Y_i^{(t)}, \Theta_i^{(t)}; \Theta, \Psi, V, \Sigma),$$

where $(Y_i^{(t)}, \Theta_i^{(t)}) \overset{iid}{\sim} [Y_i, \Theta_i|Z_i, \Theta^{(k)}, \Psi^{(k)}, V^{(k)}, \Sigma^{(k)}]$.

A.2 Stochastic EM Algorithm

At each iteration, the algorithm imputes one drawing of missing variables conditional on observable data and current estimate of the parameters and then maximizes the pseudo-complete-data likelihood function. Fundamentally, it is the same as the above MCEM algorithm with $l_{k+1} \equiv 1$.

A.3 Stochastic Approximation EM Algorithm

With a sequence of positive step sizes $\{\gamma_k\}_{k\geq 1}$, the conditional expectations (5) are updated iteratively as

$$\begin{aligned}&\hat{\mathrm{E}}^{(k+1)}[g(Y_i, \Theta_i; \Theta, \Psi, V, \Sigma)|Z_i] \\ &\quad = (1-\gamma_{k+1})\hat{\mathrm{E}}^{(k)}[g(Y_i, \Theta_i; \Theta, \Psi, V, \Sigma)|Z_i] + \gamma_{k+1}\bar{g}^{(k+1)}(Y_i, \Theta_i; \Theta, \Psi, V, \Sigma)\end{aligned}$$

where $\bar{g}^{(k+1)}(\cdot)$ is the average of $g(\cdot)$ by using the simulated missing data in $(k+1)$-st iteration. The point estimate of the parameter is computed by averaging the sequence of parameter estimates in the SAEM algorithm or seeking the converging point of the sequence. For the convergence of the point estimation, it is required that $0 \leq \gamma_k \leq 1$, $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$ but $\sum_{k=1}^{\infty} \gamma_k = \infty$.

References

1. Amemiya T (1985) Advanced econometrics. Harvard University Press, Cambridge
2. Berger J (2000) Bayesian analysis: A look at today and thoughts of tomorrow. J Am Stat Assoc 95:1269–1276
3. Bollen K (1989) Structural equations with latent variables. Wiley, New York
4. Broniatowski M, Celeux G, Diebolt J (1983) Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste. In: Diday E (ed) Data analysis and informatics, vol 3. Amsterdam, North Holland, pp 359–374

5. Cappé O, Robert C (2000) Markov chain monte carlo: 10 years and still running! J Am Stat Assoc 95:1282–1286
6. Catalano P, Ryan L (1992) Bivariate latent variable models for clustered discrete and continuous outcomes. J Am Stat Assoc 87:651–658
7. Celeux G, Diebolt J (1983) A probabilistic teacher algorithm for iterative maximum likelihood estimation. In: Bock H (ed) Classification and related methods of data analysis. Amsterdam, North Holland, pp 617–623
8. Celeux G, Diebolt J (1985) The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Comput Stat Q 2:73–82
9. Chan J, Kuk A (1997) Maximum likelihood estimation for probit-linear mixed models with correlated random effects. Biometrics 53:86–97
10. Daniels M, Pourahmadi M (2002) Bayesian analysis of covariance matrices and dynamic models for longitudinal data. Biometrika 89:553–566
11. Delyon B, Lavielle M, Moulines E (1999) Convergence of a stochastic approximation version of the EM algorithm. Ann Stat 27:94–128
12. Dempster A, Laird N, Rubin D (1977) Maximum likelihood estimation form incomplete data via the em algorithm (with discussion). J R Stat Soc Series B 39:1–38
13. Dunson D (2000) Bayesian latent variable models for clustered mixed outcomes. J R Stat Soc Series B 62:355–366
14. Dunson D, Chen Z, Harry J (2003) A bayesian approach for joint modeling of cluster size and subunit-specific outcomes. Biometrics 59:521–530
15. Edwards D (1990) Hierarchical interaction models (with discussion). J R Stat Soc Series B 52:3–20
16. Edwards D, Lauritzen S (2001) The TM algorithm for maximising a conditional likelihood function. Biometrika 88:961–972
17. Gueorguieva R, Agresti A (2001) A correlated probit model for joint modelling of clustered binary and continuous responses. J Am Stat Assoc 96:1102–1112
18. Gueorguieva R, Sanacora G (2006) Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. Stat Med 25:1307–1322
19. Haavelmo T (1943) The statistical implications of a system of simultaneous equations. Econometrica 11:1–12
20. Heckman J, MaCurdy T (1980) A life cycle model of female labour supply. Rev Econ Stud 47:47–74
21. Koster J (1996) Markov properties of non-recursive causal models. Ann Stat 24:2148–2177
22. Lauritzen S, Wermuth N (1989) Graphical models for associations between variables, some of which are qualitative and some quantitative. Ann Stat 17:31–57
23. Lin X, Ryan L, Sammel M, Zhang D, Padungtod C, Xu X (2000) A scaled linear mixed model for multiple outcomes. Biometrics 56:593–601
24. Louis T (1982) Finding the observed information matrix when using the EM algorithm. J R Stat Soc Series B 44:226–233
25. McLachlan GJ, Krishnan T (1997) The EM algorithm and extensions. Wiley, New York
26. Miglioretti D (2003) Latent transition regression for mixed outcomes. Biometrics 59:710–720
27. Price C, Kimmel C, Tyl R, Marr M (1985) The developmental toxicity of ethylene glycol in rats and mice. Toxicol Appl Pharmacol 81:113–127
28. Rochon J (1996) Analyzing bivariate repeated measures for discrete and continuous outcome variables. Biometrics 52:740–750
29. Roy J, Lin X, Ryan L (2003) Scaled marginal models for multiple continuous outcomes. Biostatistics 4:371–383
30. Sammel M, Lin X, Ryan L (1999) Multivariate linear mixed models for multiple outcomes. Stat Med 18:2479–2492
31. Wei G, Tanner M (199) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. J Am Stat Assoc 85:699–704
32. Zellner A (1962) An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. J Am Stat Assoc 57:348–368