Robustness and Complex Data Structures

Claudia Becker • Roland Fried • Sonja Kuhnt
Editors

Robustness and Complex Data Structures

Festschrift in Honour of Ursula Gather

Springer

Editors

Claudia Becker
Faculty of Law, Economics, and Business
Martin-Luther-University Halle-Wittenberg
Halle, Germany

Sonja Kuhnt
Faculty of Statistics
TU Dortmund University
Dortmund, Germany

Roland Fried
Faculty of Statistics
TU Dortmund University
Dortmund, Germany

ISBN 978-3-642-35493-9 ISBN 978-3-642-35494-6 (eBook)
DOI 10.1007/978-3-642-35494-6
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013932868

© Springer-Verlag Berlin Heidelberg 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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

Foreword

Elisabeth Noelle-Neumann, Professor of Communication Sciences at the University of Mainz and Founder of the Institut für Demoskopie Allensbach, once declared:

> *"For me, statistics is the information source of the responsible. (...) The sentence: 'with statistics it is possible to prove anything' serves only the comfortable, those who have no inclination to examine things more closely."*[1]

Examining things closely, engaging in exact analysis of circumstances as the basis for determining a course of action are what Ursula Gather is known for, and what she passes on to future generations of scholars. Be it as Professor of Mathematical Statistics and Applications in Industry at the Technical University of Dortmund, in her role, since 2008, as Rector of the TU Dortmund, or as a member of numerous leading scientific committees and institutions, she has dedicated herself to the service of academia in Germany and abroad.

In her career, Ursula Gather has combined scientific excellence with active participation in university self-administration. In doing so, she has never settled for the easy path, but has constantly searched for new insights and challenges. Her expertise, which ranges from complex statistical theory to applied research in the area of process planning in forming technology as well as online monitoring in intensive care in the medical sciences, is widely respected. Her reputation reaches far beyond Germany's borders and her research has been awarded prizes around the world.

It has been both a great pleasure and professionally enriching for me to have been fortunate enough to cooperate with her across the boundaries of our respective scientific disciplines, and I know that in this I am not alone. The success of the internationally renowned DFG Collaborative Research Centre 475 "Reduction of Complexity for Multivariate Data Structures" was due in large part to Ursula Gather's leadership over its entire running time of 12 years (1997–2009). She has also given

[1] "Statistik ist für mich das Informationsmittel der Mündigen. (...) Der Satz: 'Mit Statistik kann man alles beweisen' gilt nur für die Bequemen, die keine Lust haben, genau hinzusehen." Quoted in: Küchenhoff, Helmut (2006), 'Statistik für Kommunikationswissenschaftler', 2nd revised edition, Konstanz: UVK-Verlags-Gesellschaft, p.14.

her time and support to the DFG over many years: From 2004 until 2011, she was a member of the Review Board Mathematics, taking on the role of chairperson from 2008 to 2011. During her years on the Review Board, she took part in more than 30 meetings, contributing to decision-making process that led to recommendations on more than 1200 individual project proposals in the field of mathematics, totalling applications for a combined sum of almost 200 million. Alongside individual project proposals and applications to programmes supporting early-career researchers, as a member of the Review Board she also played an exemplary role in the selection of projects for the DFG's coordinated research programmes.

Academic quality and excellence always underpin the work of Ursula Gather. Above and beyond this, however, she possesses a clear sense of people as well as a keen understanding of the fundamental questions at hand. The list of her achievements and organizational affiliations is long; too long to reproduce in its entirety here. Nonetheless, her work as an academic manager should not go undocumented. Since her appointment as Professor of Mathematical Statistics and Applications in Industry in 1986, she has played a central role in the development of the Technical University of Dortmund, not least as Dean of the Faculty of Statistics and later Pro-Rector for Research. And, of course, as Rector of the University since 2008 she has also had a very significant impact on its development. It is not least as a result of her vision and leadership that the Technical University has come to shape the identity of Dortmund as a centre of academia and scientific research. The importance of the Technical University for the city of Dortmund, for the region and for science in Germany was also apparent during the General Assembly of the DFG in 2012, during which we enjoyed the hospitality of the TU Dortmund. Ursula Gather can be proud of what she has achieved. It will, however, be clear to everyone who knows her and has had the pleasure of working with her that she is far from the end of her achievements. I for one am happy to know that we can all look forward to many further years of working with her.

Personalities like Ursula Gather drive science forward with enthusiasm, engagement, inspiration and great personal dedication. Ursula, I would like, therefore, to express my heartfelt thanks for your work, for your close cooperation in diverse academic contexts and for your support personally over so many years. My thanks go to you as a much respected colleague and trusted counsellor, but also as a friend. Many congratulations and my best wishes on the occasion of your sixtieth birthday!

Bonn, Germany
November 2012

Matthias Kleiner
President of the German
Research Foundation

Preface

Our journey towards this Festschrift started when realizing that our teacher, mentor, and friend Ursula Gather was going to celebrate her 60th birthday soon. As a researcher, lecturer, scientific advisor, board member, reviewer, editor, Ursula has had a wide impact on Statistics in Germany and within the international community. So we came up with the idea of following the good academic tradition of dedicating a Festschrift to her. We aimed at contributions from highly recognized fellow researchers, former students and project partners from various periods of Ursula's academic career, covering a wide variety of topics from her main research interests. We received very positive responses, and all contributors were very much delighted to express their gratitude and sympathy to Ursula in this way. And here we are today, presenting this interesting collection, divided into three main topics which are representatives of her research areas.

Starting from questions on outliers and extreme value theory, Ursula's research interests spread out to cover robust methods—from Ph.D. through habilitation up to leading her own scholars to this field, including us, robust and nonparametric methods for high-dimensional data and time series—particularly within the collaborative research center SFB 475 "Reduction of Complexity in Multivariate Data Structures", up to investigating complex data structures—manifesting in projects in the research centers SFB 475 and SFB 823 "Statistical Modelling of Nonlinear Dynamic Processes".

The three parts of this book are arranged according to these general topics. All contributions aim at providing an insight into the research field by easy-to-read introductions to the various themes. In the first part, contributions range from robust estimation of location and scatter, over breakdown points, outlier definition and identification, up to robustness for non-standard multivariate data structures. The second part covers regression scenarios as well as various aspects of time series analysis like change point detection and signal extraction, robust estimation, and outlier detection. Finally, the analysis of complex data structures is treated. Support vector machines, machine learning, and data mining show the link to ideas from information science. The (lack of) relation between correlation analysis and tail dependence or diversification effects in financial crisis is clarified. Measures of statistical evidence are introduced, complex data structures are uncovered by graphical models, a data mining approach on pharmacoepidemiological databases is analyzed and meta analysis in clinical trials has to deal with complex combination of separate studies.

We are grateful to the authors for their positive response and easy cooperation at the various steps of developing the book. Without all of you, this would not have been possible. We apologize to all colleagues we did not contact as our selection is of course strongly biased by our own experiences and memories. We hope that you enjoy reading this Festschrift nonetheless. Our special thanks go to Matthias Borowski at TU Dortmund University for supporting the genesis of this work with patient help in all questions of the editing process and his invaluable support in preparing the final document, and to Alice Blanck at Springer for encouraging us to go on this wonderful adventure and for helping us finishing it. Our biggest thanks of course go to Ursula, who introduced us to these fascinating research fields and the wonderful people who have contributed to this Festschrift. Without you, Ursula, none of this would have been possible!

Halle and Dortmund, Germany
April 2013

Claudia Becker
Roland Fried
Sonja Kuhnt

Contents

Part I Univariate and Multivariate Robust Methods

Part II Regression and Time Series Analysis

Part III Complex Data Structures

Part I
Univariate and Multivariate Robust Methods

Chapter 1
Multivariate Median

Hannu Oja

1.1 Introduction

Multivariate medians are robust competitors of the mean vector in estimating the symmetry center of a multivariate distribution. Various definitions of multivariate medians have been proposed in the literature, and their properties (efficiency, equivariance, robustness, computational convenience, estimation of their accuracy, etc.) have been extensively investigated. The univariate median as well as the univariate concepts of sign and rank are based on the ordering of the univariate observations. Unfortunately, there is no natural ordering of multivariate data points. An approach utilizing L_1 objective functions is therefore often used to extend these concepts to the multivariate case. In this paper, we consider three multivariate extensions of the median, the vector of marginal medians, the spatial median, and the Oja median, based on three different multivariate L_1 objective functions, and review their statistical properties as found in the literature. For other reviews of the multivariate median, see Small (1990), Chaudhuri and Sengupta (1993), Niinimaa and Oja (1999), Dhar and Chauduri (2011).

A brief outline of the contents of this chapter is as follows. We trace the ideas in the univariate case. Therefore, in Sect. 1.2 we review the univariate concepts of sign and rank with corresponding tests and the univariate median with possible criterion functions for its definition. The first extension based on the so called Manhattan distance is the vector of marginal medians, and its properties are discussed in Sect. 1.3. The use of the Euclidean distance in Sect. 1.4 determines the spatial median and, finally in Sect. 1.5, the sum of the volumes of the simplices based on data points are used to build the objective function for the multivariate Oja median. The statistical properties of these three extensions of the median are carefully reviewed and comparisons are made between them. The chapter ends with a short conclusion in Sect. 1.7.

H. Oja (✉)
Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland
e-mail: hannu.oja@utu.fi

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_1, © Springer-Verlag Berlin Heidelberg 2013

1.2 Univariate Median

Let $\mathbf{x} = (x_1, \dots, x_n)'$ be a random sample from a univariate distribution with cumulative distribution function F. The median functional $T(F)$ and the corresponding sample statistic $T(\mathbf{x}) = T(F_n)$ can be defined in several ways. Some possible definitions for the univariate median follow.

1. The median functional is defined as
$$T(F) = \inf\left\{x : F(x) \geq \frac{1}{2}\right\}.$$
2. The median $T(F)$ maximizes the function
$$t \to \min\big\{P(x_1 \leq t), P(x_1 \geq t)\big\} = \min\big\{F(t), 1 - F(t-)\big\}.$$
3. The median $T(F)$ maximizes the function
$$t \to P\big(\min\{x_1, x_2\} \leq t \leq \max\{x_1, x_2\}\big) = 2F(t)\big(1 - F(t-)\big).$$
4. The median $T(F)$ minimizes
$$E\big(|x_1 - t|\big) \quad \text{or} \quad D(t) = E\big\{|x_1 - t| - |x_1|\big\}.$$
Note that, as $||x_1 - t| - |x_1|| \leq |t|$, the expectation in the definition of $D(t)$ always exists.
5. The median $T(F)$ solves the estimation equation
$$E\big[S(x_1 - t)\big] = 0,$$
where $S(t)$ is the *univariate sign function*
$$S(t) = \begin{cases} +1, & \text{if } t > 0, \\ 0, & \text{if } t = 0, \\ -1, & \text{if } t < 0. \end{cases}$$

Different definitions of the population median $T(F)$ listed above all yield the same unique value μ for a distribution F with a bounded and continuous density $f(\mu)$ at μ. For the objective function $D(t)$, it is then true that
$$D(t) = D(\mu) + \frac{\delta}{2}(t - \mu)^2 + o\big((t - \mu)^2\big) \quad \text{with } \delta = 2f(\mu).$$

The sample median $\hat{\mu}$ is associated with the *univariate sign test* based on the sign function $S(t)$. Starting from the univariate sign function, the *univariate (centered) rank function* is defined as
$$\hat{R}(t) = \frac{1}{n}\sum_{i=1}^{n} S(t - x_i).$$

Note that $\hat{R}(t) \in [-1, 1]$ and that the estimating equation for the sample median is $\hat{R}(\hat{\mu}) = 0$. The sign test statistic for testing the null hypothesis $H_0 : \mu = 0$ is

$\hat{R}(0)$. The test statistic is strictly and asymptotically distribution-free, as for the true median μ,

$$n\frac{\hat{R}(\mu)+1}{2} \sim \mathrm{Bin}\left(n, \frac{1}{2}\right), \quad \text{and} \quad \sqrt{n}\hat{R}(\mu) \to_d N(0,1).$$

One can also show that

$$\hat{\mu} = \mu + \delta^{-1}\hat{R}(\mu) + o_P\left(n^{-1/2}\right),$$

where $\delta = 2f(\mu)$, and, consequently,

$$\sqrt{n}(\hat{\mu} - \mu) \to_d N_p\left(0, \delta^{-2}\right).$$

Computation When applied to the sample cdf F_n, different definitions above yield different and not necessarily unique solutions. The sample median $\hat{\mu}$, which is an estimate of the population median $T(F) = \mu$, is then usually defined as follows. First, let $x_{(1)}, \dots, x_{(n)}$ be the ordered observations. (Note that in the multivariate case there is no natural ordering of the data points.) The sample median is then

$$\hat{\mu} = \frac{x_{[(n+1)/2]} + x_{[(n+2)/2]}}{2},$$

where $[t]$ denotes the integer part of t.

Robustness It is well known that the median is a highly robust estimate with the asymptotic breakdown point $1/2$ and the bounded influence function $\mathrm{IF}(x; T, F) = \delta^{-1}S(x - T(F))$.

Asymptotic Efficiency If the distribution F has a finite second moment σ^2, then the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, that estimates the population mean $\mu = E(x_i)$, has a limiting normal distribution, and

$$\sqrt{n}(\bar{x} - \mu) \to N\left(0, \sigma^2\right).$$

For symmetric F, the asymptotic relative efficiency (ARE) between the sample median and sample mean is then defined as the ratio of the limiting variances

$$\mathrm{ARE} = 4f^2(\mu)\sigma^2.$$

If F is the normal distribution $N(\mu, \sigma^2)$, this ARE $= 0.64$ is small. However, for heavy-tailed distributions, the asymptotic efficiency of the median is better; AREs for a t-distribution with 3 degrees of freedom and for a Laplace distribution are, for example, 1.62 and 2.

Estimation of the Variance of the Estimate Estimation of $\delta = 2f(\mu)$ from the data is difficult. For a discussion, see Example 1.5.5 in Hettmansperger and McKean (1998) and Oja (1999). It is, however, remarkable that by inverting the sign test, it is possible to obtain strictly distribution-free confidence intervals for μ. This follows as, for a continuous distribution F,

$$P(x_{(i)} < \mu < x_{(n+1-i)}) = P\left(i \le \frac{nR(\mu)+1}{2} \le n-i\right) = \sum_{j=i}^{n-i}\binom{n}{j}2^{-n}.$$

Equivariance For a location functional, one hopes that the functional is *equivariant under linear transformations*, that is,

$$T(F_{ax+b}) = aT(F_x) + b, \quad \text{for all } a \text{ and } b.$$

This is true for the median functional in the family of distributions with bounded and continuous derivative at the median. Note also that the median is in fact equivariant under much larger sets of transformations. If $g(x)$ is any strictly monotone function, then $T(F_{g(x)}) = g(T(F_x))$.

Location M-estimates The sample median is a member of the family of M-estimates. Assume for a moment that $\mathbf{x} = (x_1, \ldots, x_n)'$ is a random sample from a continuous distribution with density function $f(x - \mu)$, where $f(x)$ is symmetric around zero. Assume also that the derivative function $f'(x)$ exists, and write $l(x) = f'(x)/f(x)$ for a location score function. The so called *location M-functionals* $T(F)$ are often defined as μ that minimizes

$$D(t) = E\big(\rho(x - t)\big)$$

with some function $\rho(t)$, or solves the estimating equation

$$R(\mu) = E\big(\psi(x - \mu)\big) = 0,$$

for an odd smooth function $\psi(t) = \rho'(t)$. The so called M-test statistic for testing $H_0 : \mu = 0$ satisfies

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(x_i) \to_d N(0, \omega) \quad \text{with } \omega = E\big(\psi^2(x - \mu)\big).$$

The M-estimate $\hat{\mu} = T(F_n)$ solves the estimating equation

$$\frac{1}{n} \sum_{i=1}^{n} \psi(x_i - \hat{\mu}) = 0$$

and, under general assumptions,

$$\sqrt{n}(\hat{\mu} - \mu) \to_d N_p\big(\mathbf{0}, \omega/\delta^2\big),$$

where the constant δ is, depending on the properties of function $\rho(t)$ and $\psi(t)$ and density $f(z)$ of $z = x_i - \mu$, given by

$$\delta = D''(\mu), \quad \text{or} \quad \delta = R'(\mu),$$

or

$$\delta = E\big(\psi'(z)\big), \quad \text{or} \quad \delta = E\big(\psi(z)l(z)\big),$$

or

$$\delta = \int \rho(z) f''(z), \quad \text{or} \quad \delta = \int \psi(z) f'(z).$$

Note that the choice $\psi(x) = l(x)$ yields the maximum likelihood estimate with the smallest possible limiting variance. The mean and median are the ML-estimates for the normal distribution ($\psi(t) = t$) and for the double-exponential (Laplace) distribution ($\psi(t) = S(t)$), respectively.

Other Families of Location Estimates Note also that the median is also a limiting case in the set of *trimmed means*

$$T_\alpha(F) = E\big(x \mid q_{F,\alpha} \le x \le q_{F,1-\alpha}\big),$$

where $q_{F,\alpha}$ is the α-quantile of F satisfying $F(q_{F,\alpha}) = \alpha$. The so called L_α *functionals* minimize

$$E\big(|x_i - t|^\alpha\big), \quad 1 \le \alpha \le 2,$$

with the mean ($\alpha = 2$) and the median ($\alpha = 1$) as special cases.

1.3 Vector of Marginal Medians

Our first extension of the median to the multivariate case is straightforward: It is simply the vector of marginal medians. Let now $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ be a random sample from a p-variate distribution with cumulative distribution function F, and assume that the p marginal distribution have bounded densities $f_1(\mu_1), \ldots, f_p(\mu_p)$ at the uniquely defined marginal medians $\mu_1, \ldots, \mu_p$. Write $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ for the vector of marginal medians.

The vector of marginal sample medians $\mathbf{T}(\mathbf{X})$ minimizes the criterion function which is the sum of componentwise distances (Manhattan distance)

$$D_n(\mathbf{t}) = \frac{1}{n}\sum_{i=1}^{n}\big\{\big(|x_{i1} - t_1| + \cdots + |x_{ip} - t_p|\big) - \big(|x_{i1}| + \cdots + |x_{ip}|\big)\big\}.$$

The corresponding population functional $T(F)$ for the vector of population medians then minimizes

$$D(\mathbf{t}) = E\big\{\big(|x_1 - t_1| + \cdots + |x_p - t_p|\big) - \big(|x_1| + \cdots + |x_p|\big)\big\}.$$

Now we obtain

$$D(\mathbf{t}) = D(\mu) + \frac{1}{2}(\mathbf{t} - \mu)'\Delta(\mathbf{t} - \mu) + o\big(\|\mathbf{t} - \mu\|^2\big),$$

where Δ is a diagonal matrix with diagonal elements $2f_1(\mu_1), \ldots, 2f_p(\mu_p)$.

Multivariate sign and rank functions are now given as

$$\mathbf{S}(\mathbf{t}) = \begin{pmatrix} S(t_1) \\ \ldots \\ S(t_p) \end{pmatrix} \quad \text{and} \quad \hat{\mathbf{R}}(\mathbf{t}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{S}(\mathbf{t} - \mathbf{x}_i),$$

where $S(t)$ is the univariate sign function. Note that $\hat{\mathbf{R}}(\mathbf{t}) \in [-1, 1]^p$. The multivariate sign test for testing the null hypothesis $H_0 : \mu = \mathbf{0}$ is based on $\hat{\mathbf{R}}(\mathbf{0})$. The marginal distributions of $\hat{\mathbf{R}}(\boldsymbol{\mu})$ are distribution-free but, unfortunately, the joint distribution of the components of $\hat{\mathbf{R}}(\boldsymbol{\mu})$ depends on the dependence structure of the components of $\mathbf{x}_i$, and, consequently,

$$\sqrt{n}\hat{\mathbf{R}}(\mu) \to_d N_p(\mathbf{0}, \Omega),$$

where $\Omega = \mathrm{Cov}(\mathbf{S}(\mathbf{x}-\mu))$. As again,

$$\hat{\mu} = \mu + \Delta^{-1}\hat{\mathbf{R}}(\mu) + \mathbf{o}_P\big(n^{-1/2}\big),$$

we get

$$\sqrt{n}(\hat{\mu}-\mu) \rightarrow_d N_p\big(\mathbf{0}, \Delta^{-1}\Omega\Delta^{-1}\big).$$

For the estimate, and its properties See, for example, Puri and Sen (1971), Babu and Rao (1988). Some important properties of the spatial median are listed below.

Computation of the Estimate As in the univariate case.

Robustness of the Estimate As in the univariate case, this multivariate extension of the median is highly robust with the asymptotic breakdown point 1/2 and the influence function is bounded, $\mathrm{IF}(\mathbf{x}; \mathbf{T}, F) = \Delta^{-1}\mathbf{S}(\mathbf{x} - \mathbf{T}(F))$ where $\mathbf{S}(\mathbf{t})$ is the vector of marginal sign functions.

Asymptotic Efficiency of the Estimate If the distribution F has a covariance matrix Σ (with finite second moments), then the sample mean vector $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$, a natural estimate of the population mean vector $\mu = E(x_i)$, has a limiting normal distribution, and

$$\sqrt{n}(\bar{\mathbf{x}}-\mu) \rightarrow N_p(\mathbf{0}, \Sigma).$$

The asymptotic relative efficiency (ARE) between the vector of sample medians and the sample mean vector, if they estimate the same population value μ, is defined as

$$\mathrm{ARE} = \left(\frac{|\Sigma|}{|\Delta^{-1}\Omega\Delta^{-1}|}\right)^{1/p}.$$

The ARE thus compares the geometrical means of the eigenvalues of the limiting covariance matrices. The comparison is, however, fair only for affine equivariant estimates and the vector of sample medians is not affine equivariant, see below. In the case of the spherical normal distribution $N_p(\mu, \sigma^2\mathbf{I}_p)$, the ARE between the vector of sample medians and the sample mean vector is as in the univariate case and therefore does not depend on the dimension p. For dependent observations, the efficiency of the median vector may be much smaller.

Estimation of the Covariance Matrix of the Estimate One easily finds

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{S}(\mathbf{x}_i-\hat{\mu})\mathbf{S}(\mathbf{x}_i-\hat{\mu})^T\big)$$

but the estimation of Δ, i.e. the estimation of the diagonal elements $2f_1(\mu_1), \ldots, 2f_p(\mu_p)$, is as difficult as in the univariate case.

Affine Equivariance of the Estimate The vector of marginal medians is not affine equivariant: For a multivariate location functional $\mathbf{T}(F)$, it is often expected that $\mathbf{T}(F)$ is *affine equivariant*, that is,

$$\mathbf{T}(F_{\mathbf{Ax}+\mathbf{b}}) = \mathbf{AT}(F_{\mathbf{x}}) + \mathbf{b}, \quad \text{for all full-rank } p \times p \text{ matrices } \mathbf{A} \text{ and } p\text{-vectors } \mathbf{b}.$$

The vector of marginal medians is not affine equivariant as the condition is true only if $\mathbf{A}$ is a diagonal matrix with non-zero diagonal elements.

Transformation–Retransformation (TR) Estimate An affine equivariant version of the vector of marginal medians is found using the so called transformation–retransformation (TR) technique. A $p \times p$-matrix valued functional $\mathbf{G}(F)$ is called an invariant coordinate system (ICS) functional if

$$\mathbf{G}(F_{\mathbf{Ax}+\mathbf{b}}) = \mathbf{G}(F_{\mathbf{x}})\mathbf{A}^{-1}, \quad \text{for all full-rank } p \times p \text{ matrices } \mathbf{A} \text{ and } p\text{-vectors } \mathbf{b}.$$

Then the *transformation–retransformation (TR) median* functional is defined as

$$\mathbf{T}_{\mathrm{TR}}(F_{\mathbf{x}}) = \mathbf{G}(F_{\mathbf{x}})^{-1}\mathbf{T}(F_{\mathbf{G}(F_{\mathbf{x}})\mathbf{x}}).$$

For the concept of the TR median, see Chakraborty and Chaudhuri (1998). For different ICS transformations, we refer to Tyler et al. (2009), Ilmonen et al. (2012).

1.4 Spatial Median

The so-called spatial median $\mathbf{T}(\mathbf{X})$ minimizes the criterion function $\sum_i \|\mathbf{x}_i - \mathbf{t}\|$, or

$$D_n(\mathbf{t}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\|\mathbf{x}_i - \mathbf{t}\| - \|\mathbf{x}_i\|\right\},$$

where $\|\mathbf{t}\| = (t_1^2 + \cdots + t_p^2)^{1/2}$ denotes the Euclidean norm. The corresponding functional, the spatial median $T(F)$, minimizes

$$D(\mathbf{t}) = E_F\left\{\|\mathbf{x} - \mathbf{t}\| - \|\mathbf{x}\|\right\}.$$

For the asymptotic results we need the assumptions

1. The spatial median μ minimizing $D(\mathbf{t})$ is unique.
2. The distribution F_x has a bounded and continuous density at μ.

Again,

$$D(\mathbf{t}) = D(\mu) + \frac{1}{2}(\mathbf{t} - \mu)'\Delta(\mathbf{t} - \mu) + o\left(\|\mathbf{t} - \mu\|^2\right),$$

where now

$$\Delta = E\left(\frac{1}{\|\mathbf{x} - \mu\|}\left[\mathbf{I}_p - \frac{(\mathbf{x} - \mu)(\mathbf{x} - \mu)'}{\|\mathbf{x} - \mu\|^2}\right]\right).$$

The assumptions above guarantee that this expectation exists.

Multivariate spatial sign and centered rank functions are now given as

$$\mathbf{S}(\mathbf{t}) = \begin{cases} \frac{\mathbf{t}}{\|\mathbf{t}\|}, & \text{if } \mathbf{t} \neq \mathbf{0}, \\ \mathbf{0}, & \text{if } \mathbf{t} = \mathbf{0} \end{cases}$$

and

$$\hat{\mathbf{R}}(\mathbf{t}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{S}(\mathbf{t} - \mathbf{x}_i).$$

Note that the spatial sign $\mathbf{S}(\mathbf{t})$ is just a unit vector in the direction of $\mathbf{t}$, $\mathbf{t} \neq 0$. The centered rank $\hat{\mathbf{R}}(\mathbf{t})$ is lying in the unit p-ball $\mathcal{B}^p$.

The spatial sign test statistic for testing $H_0 : \mu = \mathbf{0}$ is $\mathbf{R}(\mathbf{0})$ and its limiting null distribution is given by

$$\sqrt{n}\hat{\mathbf{R}}(\boldsymbol{\mu}) \rightarrow_d N_p(\mathbf{0}, \Omega),$$

where

$$\Omega = E\left(\frac{(\mathbf{x} - \mu)(\mathbf{x} - \mu)'}{\|\mathbf{x} - \mu\|^2} \right).$$

Again,

$$\hat{\mu} = \mu + \Delta^{-1}\hat{\mathbf{R}}(\mu) + \mathbf{o}_P\left(n^{-1/2}\right),$$

and we obtain

$$\sqrt{n}(\hat{\mu} - \mu) \rightarrow_d N_p\left(0, \Delta^{-1}\Omega\Delta^{-1}\right).$$

For the properties of the estimate we refer to Oja (2010), Möttönen et al. (2010).

Computation of the Estimate The spatial median is unique if the data fall in on at least two-dimensional space. The so called Weisfeld algorithm for the computation of the spatial median has an iteration step

$$\mu \leftarrow \mu + \left[\frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i - \mu\|^{-1} \right]^{-1} \mathbf{R}(\mu).$$

The algorithm may fail sometimes but a modified algorithm by Vardi and Zhang (2000) converges fast and monotonically. The estimate with estimated covariance matrix can be obtained using the R package MNM, see Nordhausen and Oja (2011).

Robustness of the Estimate The spatial median is highly robust with the asymptotic breakdown point 1/2. The influence function is bounded, $\text{IF}(\mathbf{x}; \mathbf{T}, F) = \Delta^{-1}\mathbf{S}(\mathbf{x} - \mathbf{T}(F))$ where $\mathbf{S}(\mathbf{t})$ is the spatial sign function.

Asymptotic Efficiency of the Estimate If the covariance matrix Σ exists, then the asymptotic relative efficiency (ARE) between the spatial median and the mean vector, if they estimate the same population value μ, is

$$\text{ARE} = \left(\frac{|\Sigma|}{|\Delta^{-1}\Omega\Delta^{-1}|} \right)^{1/p}.$$

In the case of a p-variate spherical distribution of $\mathbf{x}$, $p > 1$, this ARE reduces to

$$\text{ARE}_p = \left(\frac{p-1}{p}\right)^2 E\big(\|\mathbf{x}\|^2\big) E^2\big(\|\mathbf{x}\|^{-1}\big).$$

In the p-variate spherical normal case, one then gets, for example,

$$\text{ARE}_2 = 0.785, \quad \text{ARE}_3 = 0.849, \quad \text{ARE}_6 = 0.920, \quad \text{and} \quad \text{ARE}_{10} = 0.951,$$

and the efficiency goes to 1 as $p \to \infty$. For heavy-tailed distributions, the spatial median outperforms the sample mean vector.

Estimation of the Covariance Matrix of the Estimate In this case, one easily finds an estimate for the approximate covariance matrix

$$\frac{1}{n}\Delta^{-1}\Omega\Delta^{-1}$$

using

$$\hat{\Delta} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\|\mathbf{x}_i - \hat{\mu}\|}\left[\mathbf{I}_p - \frac{(\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})'}{\|\mathbf{x}_i - \hat{\mu}\|^2}\right]\right)$$

and

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\frac{(\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})'}{\|\mathbf{x}_i - \hat{\mu}\|^2}.$$

Estimation of the covariance matrix of the spatial median is implemented in the R package MNM.

Affine Equivariance of the Estimate The spatial median is not affine equivariant as

$$\mathbf{T}(F_{\mathbf{Ax+b}}) = \mathbf{AT}(F_{\mathbf{x}}) + \mathbf{b}$$

is true only for orthogonal matrices $\mathbf{A}$.

Transformation–Retransformation (TR) Estimate An affine equivariant transformation retransformation (TR) spatial median is found as follows. Let $\mathbf{S}(F)$ be a scatter functional, and find a $p \times p$-matrix valued functional $\mathbf{G}(F) = \mathbf{S}^{-1/2}(F)$ such that

$$\mathbf{G}(F)\mathbf{S}(F)\mathbf{G}(F)' = \mathbf{I}_p.$$

Note that $\mathbf{G}(F)$ is not necessarily an invariant coordinate functional. Then the *transformation–retransformation (TR) median* is

$$\mathbf{T}_{\text{TR}}(F_{\mathbf{x}}) = \mathbf{G}(F_{\mathbf{x}})^{-1}\mathbf{T}(F_{\mathbf{G}(F_{\mathbf{x}})\mathbf{x}}),$$

see Chakraborty et al. (1998), Ilmonen et al. (2012). The TR median that combines the spatial median and Tyler's scatter matrix was proposed in Hettmansperger and Randles (2002) and is called the Hettmansperger–Randles median. It can be computed using the R package MNM.

1.5 Oja Median

Let again $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ be a random sample from a p-variate distribution with cumulative distribution function F. The volume of the p-variate simplex determined by $p+1$ vertices $\mathbf{t}_1, \ldots, \mathbf{t}_{p+1}$ is

$$V(\mathbf{t}_1, \ldots, \mathbf{t}_{p+1}) = \frac{1}{p!} \left| \det \begin{pmatrix} 1 & \ldots & 1 \\ \mathbf{t}_1 & \ldots & \mathbf{t}_{p+1} \end{pmatrix} \right|.$$

Note that, in the univariate case $V(t_1, t_2)$ is the length of the interval with endpoints in t_1 and t_2, in the bivariate case $V(\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3)$ is the area of the triangle with corners at $\mathbf{t}_1$, $\mathbf{t}_2$, and $\mathbf{t}_3$, and so on.

The so called Oja median (estimate) $\mathbf{T}(\mathbf{X})$ minimizes the objective function

$$D_n(\mathbf{t}) = \binom{n}{p}^{-1} \sum_{i_1 < \cdots < i_p} V(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_p}, \mathbf{t}).$$

The corresponding functional $\mathbf{T}(F)$ minimizes

$$D(\mathbf{t}) = E_F\big\{V(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_p}, \mathbf{t})\big\}.$$

Note that the definition of this functional requires the existence of first moments. The vector of marginal medians and the spatial median do not need that assumptions. For the asymptotic results, we also need the assumptions that (i) the Oja median μ minimizing $D(\mathbf{t})$ is unique, and that (ii) the second moments exist. One can again write

$$D(\mathbf{t}) = D(\mu) + \frac{1}{2}(\mathbf{t} - \mu)'\Delta(\mathbf{t} - \mu) + o\big(\|\mathbf{t} - \mu\|^2\big) \quad \text{with } \Delta = \frac{\partial^2}{\partial \mathbf{t} \partial \mathbf{t}'} D(\mathbf{t})\bigg|_{\mathbf{t}=\mu}.$$

Consider next the corresponding multivariate sign and rank concept. To simplify the notations, write

$$\mathcal{Q} = \big\{q = (i_1, \ldots, i_{p-1}) : 1 \le i_1 < \cdots < i_{p-1} \le n\big\}$$

and

$$\mathcal{P} = \big\{p = (i_1, \ldots, i_p) : 1 \le i_1 < \cdots < i_p \le n\big\}.$$

In the following, $q \in \mathcal{Q}$ and $p \in \mathcal{P}$ are used as indices for $(p-1)$ and p-subsets of observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Next define $\mathbf{e}_q$, d_{0p} and $\mathbf{d}_p$ through the equations

$$\det(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_{p-1}}, \mathbf{x}) = \mathbf{e}_q'\mathbf{x} \quad \text{and} \quad \det \begin{pmatrix} 1 & \ldots & 1 & 1 \\ \mathbf{x}_{i_1} & \ldots & \mathbf{x}_{i_p} & \mathbf{x} \end{pmatrix} = d_{0p} + \mathbf{d}_p'\mathbf{x}.$$

The sign and rank functions are then defined as

$$\hat{\mathbf{S}}(\mathbf{t}) = \binom{n}{q}^{-1} \sum_{q \in \mathcal{Q}} \text{sign}\big(\mathbf{e}_q'\mathbf{t}\big)\mathbf{e}_q \quad \text{and} \quad \hat{\mathbf{R}}(\mathbf{t}) = \binom{n}{p}^{-1} \sum_{p \in \mathcal{P}} \text{sign}\big(d_{0p} + \mathbf{d}_p'\mathbf{t}\big)\mathbf{d}_p.$$

The population (theoretical) sign and rank functions are then

$$\mathbf{S}(\mathbf{t}) = E\big(\text{sign}\big(\mathbf{e}_q'\mathbf{t}\big)\mathbf{e}_q\big) \quad \text{and} \quad \mathbf{R}(\mathbf{t}) = E\big(\text{sign}\big(d_{0p} + \mathbf{d}_p'\mathbf{t}\big)\mathbf{d}_p\big),$$

respectively.

The sample Oja median then solves the estimation equation $\hat{\mathbf{R}}(\hat{\mu}) = \mathbf{0}$. The sign test statistic for testing the null hypothesis $H_0 : \mu = 0$ is

$$\mathbf{T}_n = \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{S}}(\mathbf{x}_i), \quad \text{which is proportional to } \hat{\mathbf{R}}(\mathbf{0}).$$

Under the null hypothesis and under some weak assumptions,

$$\sqrt{n}\mathbf{T}_n \to_d N_p(\mathbf{0}, \Omega) \quad \text{with } \Omega = E\big(\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})'\big).$$

Again, for $\mu = \mathbf{0}$,

$$\hat{\mu} = \Delta^{-1}\mathbf{T}_n + \mathbf{o}_P\big(n^{-1/2}\big),$$

and we obtain, for true value of μ,

$$\sqrt{n}(\hat{\mu} - \mu) \to_d N_p\big(0, \Delta^{-1}\Omega\Delta^{-1}\big).$$

For the Oja median and its basic properties, see Oja (1983, 1999). For the asymptotics, we refer to Arcones et al. (1994), Shen (2008).

Computation of the Estimate The computation of the Oja median is a demanding task. The Oja median may be computed using the R-package OjaNP. See also Ronkainen et al. (2002).

Robustness of the Estimate The breakdown point of the Oja median is zero. However, if the first moments exist, then the influence function is bounded.

Asymptotic Efficiency of the Estimate In the spherical case the asymptotic efficiencies of the Oja median and the spatial median are the same (if the second moments exist); the Oja median outperforms the spatial median in the elliptic case (if the second moments exist).

Estimation of the Covariance Matrix of the Estimate See Nadar et al. (2003).

Affine Equivariance of the Estimate Unlike the vector of marginal medians and the spatial median, the Oja median is affine equivariant.

1.6 Other Medians

If in the univariate case, x_1 and x_2 are two independent observations from F, the univariate median of F could also be defined as a point μ with highest probability $P(\min\{x_1, x_2\} \le \mu \le \max\{x_1, x_2\})$. The sample median is the point lying in the largest number of data based intervals (univariate simplices). The multivariate Liu median (or simplicial depth median) of p-variate data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is then the point lying in the largest number of data based p-variate simplices. See Liu (1990) for the definition and some basic properties. For the asymptotics of the Liu median, see Arcones et al. (1994). In the bivariate normal case, the Liu median and the Oja median has the same asymptotic efficiency (if the second moments exist): The Liu median is affine equivariant with a limiting breakdown point below $1/(p+2)$.

The multivariate half-space depth function is a natural multivariate extension of the univariate median criterion function $\min\{P(x_1 \leq \mu), P(x_1 \geq \mu)\}$. The so called half-space median or the Tukey median maximizes the half space depth function, see Donoho and Gasko (1992). The half-space median is more robust than the Oja median or Liu median in the sense that its breakdown point is $1/3$. For the asymptotics, see Masse (2002).

1.7 Conclusions

In this chapter, we compared different extensions of multivariate medians. The choice of the median for a practical data analysis strongly depends on the application. The vector of marginal medians and the spatial median are highly robust but they are not affine equivariant. The efficiency of the vector of marginal medians is poor as compared to the spatial median and the Oja median. The spatial median and its affine equivariant version, the Hettmansperger–Randles median, are the only medians for which an estimate of the covariance matrix can be computed in practice with the R package MNM. This allows statistical inference with confidence ellipsoids, for example. The author's favorite median is therefore the Hettmansperger–Randles median, see Möttönen et al. (2010). For other estimators of multivariate location, see the contribution by Rousseeuw and Hubert, Chap. 4.

References

Arcones, M. A., Chen, Z., & Gine, E. (1994). Estimators related to U-processes with applications to multivariate medians: asymptotic normality. *The Annals of Statistics*, *22*, 1460–1477.

Babu, G. J., & Rao, C. R. (1988). Joint asymptotic distribution of marginal quantile functions in samples from multivariate population. *Journal of Multivariate Analysis*, *27*, 15–23.

Chakraborty, B., & Chaudhuri, P. (1998). On an adaptive transformation retransformation estimate of multivariate location. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *60*, 145–157.

Chakraborty, B., Chaudhuri, P., & Oja, H. (1998). Operating transformation retarnsformation on spatial median and angle test. *Statistica Sinica*, *8*, 767–784.

Chaudhuri, P., & Sengupta, D. (1993). Sign tests in multidimension: Inference based on the geometry of data cloud. *Journal of the American Statistical Association*, *88*, 1363–1370.

Dhar, S. S., & Chauduri, P. (2011). On the statistical efficiency of robust estimators of multivariate location. *Statistical Methodology*, *8*, 113–128.

Donoho, D. L., & Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, *20*, 1803–1827.

Hettmansperger, T. P., & McKean, J. W. (1998). *Robust nonparametric statistical methods*. London: Arnold.

Hettmansperger, T. P., & Randles, R. (2002). A practical affine equivariant multivariate median. *Biometrika*, *89*, 851–860.

Ilmonen, P., Oja, H., & Serfling, R. (2012). On invariant coordinate system (ICS) functionals. *International Statistical Review*, *80*, 93–110.

Liu, R. Y. (1990). On the notion of data depth based upon random simplices. *The Annals of Statistics*, *18*, 405–414.

Masse, J. C. (2002). Asymptotics for the Tukey median. *Journal of Multivariate Analysis*, *81*, 286–300.

Möttönen, J., Nordhausen, K., & Oja, H. (2010). Asymptotic theory of the spatial median. In *IMS collections: Vol. 7. Festschrift in honor of professor Jana Jureckova* (pp. 182–193).

Nadar, M., Hettmansperger, T. P., & Oja, H. (2003). The asymptotic variance of the Oja median. *Statistics & Probability Letters*, *64*, 431–442.

Niinimaa, A., & Oja, H. (1999). Multivariate median. In S. Kotz, N. L. Johnson, & C. P. Read (Eds.), *Encyclopedia of statistical sciences* (Vol. 3). New York: Wiley.

Nordhausen, K., & Oja, H. (2011). Multivariate L1 methods: the package MNM. *Journal of Statistical Software*, *43*, 1–28.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, *1*, 327–332.

Oja, H. (1999). Affine invariant multivariate sign and rank tests and corresponding estimates: a review. *Scandinavian Journal of Statistics*, *26*, 319–343.

Oja, H. (2010). *Multivariate nonparametric methods with R. An approach based on spatial signs and ranks*. New York: Springer.

Puri, M. L., & Sen, P. K. (1971). *Nonparametric methods in multivariate analysis*. New York: Wiley.

Ronkainen, T., Oja, H., & Orponen, P. (2002). Computation of the multivariate Oja median. In R. Dutter, P. Filzmoser, U. Gather, & P. J. Rousseeuw (Eds.), *Developments in robust statistics* (pp. 344–359). Heidelberg: Springer.

Shen, G. (2008). Asymptotics of the Oja median estimate. *Statistics & Probability Letters*, *78*, 2137–2141.

Small, G. (1990). A survey of multidimensional medians. *International Statistical Review*, *58*, 263–277.

Tyler, D., Critchley, F., Dumbgen, L., & Oja, H. (2009). Invariant coordinate selection. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *71*, 549–592.

Vardi, Y., & Zhang, C.-H. (2000). The multivariate L_1 median and associated data depth. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 1423–1426.

Chapter 2
Depth Statistics

Karl Mosler

2.1 Introduction

In 1975, John Tukey proposed a multivariate median which is the 'deepest' point in a given data cloud in $\mathbb{R}^d$ (Tukey 1975). In measuring the depth of an arbitrary point z with respect to the data, Donoho and Gasko (1992) considered hyperplanes through z and determined its 'depth' by the smallest portion of data that are separated by such a hyperplane. Since then, this idea has proved extremely fruitful. A rich statistical methodology has developed that is based on data depth and, more general, nonparametric depth statistics. General notions of data depth have been introduced as well as many special ones. These notions vary regarding their computability and robustness and their sensitivity to reflect asymmetric shapes of the data. According to their different properties they fit to particular applications. The upper level sets of a depth statistic provide a family of set-valued statistics, named *depth-trimmed* or *central regions*. They describe the distribution regarding its location, scale and shape. The most central region serves as a *median*; see also the contribution by Oja, Chap. 1. The notion of depth has been extended from data clouds, that is empirical distributions, to general probability distributions on $\mathbb{R}^d$, thus allowing for laws of large numbers and consistency results. It has also been extended from d-variate data to data in functional spaces. The present chapter surveys the theory and methodology of depth statistics.

Recent reviews on data depth are given in Cascos (2009) and Serfling (2006). Liu et al. (2006) collects theoretical as well as applied work. More on the theory of depth functions and many details are found in Zuo and Serfling (2000) and the monograph by Mosler (2002).

The depth of a data point is reversely related to its *outlyingness*, and the depth-trimmed regions can be seen as multivariate set-valued *quantiles*. To illustrate the

K. Mosler (✉)
Universität zu Köln, Albertus-Magnus-Platz, 50923 Köln, Germany
e-mail: mosler@statistik.uni-koeln.de

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_2, © Springer-Verlag Berlin Heidelberg 2013

Table 2.1 General government gross debt (% of GDP) and unemployment rate of the EU-27 countries in 2011 (Source: EUROSTAT)

Country	Debt %	Unempl. %	Country	Debt %	Unempl. %
Belgium	98.0	7.2	Luxembourg	18.2	4.9
Bulgaria	16.3	11.3	Hungary	80.6	10.9
Czech Republic	41.2	6.7	Malta	72.0	6.5
Denmark	46.5	7.6	Netherlands	65.2	4.4
Germany	81.2	5.9	Austria	72.2	4.2
Estonia	6.0	12.5	Poland	56.3	9.7
Ireland	108.2	14.4	Portugal	107.8	12.9
Greece	165.3	17.7	Romania	33.3	7.4
Spain	68.5	21.7	Slovenia	47.6	8.2
France	85.8	9.6	Slovakia	43.3	13.6
Italy	120.1	8.4	Finland	48.6	7.8
Cyprus	71.6	7.9	Sweden	38.4	7.5
Latvia	42.6	16.2	United Kingdom	85.7	8.0
Lithuania	38.5	15.4			

notions, we consider bivariate data from the EU-27 countries regarding unemployment rate and general government debt in percent of the GDP (Table 2.1). In what follows, we are interested which countries belong to a central, rather homogeneous group and which have to be regarded as, in some sense, outlying.

Section 2.2 introduces general depth statistics and the notions related to it. In Sect. 2.3, various depths for d-variate data are surveyed: multivariate depths based on distances, weighted means, halfspaces or simplices. Section 2.4 provides an approach to depth for functional data, while Sect. 2.5 treats computational issues. Section 2.6 concludes with remarks on applications.

2.2 Basic Concepts

In this section, the basic concepts of depth statistics are introduced, together with several related notions. First, we provide a general notion of depth functions, which relies on a set of desirable properties; then a few variants of the properties are discussed (Sect. 2.2.1). A depth function induces an outlyingness function and a family of central regions (Sect. 2.2.2). Further, a stochastic ordering and a probability metric are generated (Sect. 2.2.3).

2.2.1 Postulates on a Depth Statistic

Let E be a Banach space, $\mathcal{B}$ its Borel sets in E, and $\mathcal{P}$ a set of probability distributions on $\mathcal{B}$. To start with and in the spirit of Tukey's approach to data analysis, we may regard $\mathcal{P}$ as the class of empirical distributions giving equal probabilities $\frac{1}{n}$ to n, not necessarily different, data points in $E = \mathbb{R}^d$.

A *depth function* is a function $D : E \times \mathcal{P} \to [0, 1]$, $(z, P) \mapsto D(z \mid P)$, that satisfies the restrictions (or 'postulates') **D1** to **D5** given below. For easier notation, we write $D(z \mid X)$ in place of $D(z \mid P)$, where X is an arbitrary random variable distributed as P. For $z \in E$, $P \in \mathcal{P}$, and any random variable X having distribution P it holds:

- **D1** *Translation invariant*: $D(z + b \mid X + b) = D(z \mid X)$ for all $b \in E$.
- **D2** *Linear invariant*: $D(Az \mid AX) = D(z \mid X)$ for every bijective linear transformation $A : E \to E$.
- **D3** *Null at infinity*: $\lim_{\|z\|\to\infty} D(z \mid X) = 0$.
- **D4** *Monotone on rays*: If a point z^* has maximal depth, that is $D(z^* \mid X) = \max_{z\in E} D(z \mid X)$, then for any r in the unit sphere of E the function $\alpha \mapsto D(z^* + \alpha r \mid X)$ decreases, in the weak sense, with $\alpha > 0$.
- **D5** *Upper semicontinuous*: The upper level sets $D_\alpha(X) = \{z \in E : D(z \mid X) \geq \alpha\}$ are closed for all α.

D1 and **D2** state that a depth function is *affine invariant*. **D3** and **D4** mean that the level sets D_α, $\alpha > 0$, are bounded and starshaped about z^*. If there is a point of maximum depth, this depth will w.l.o.g. be set to 1. **D5** is a useful technical restriction. An immediate consequence of restriction **D4** is the following proposition.

Proposition 2.1 *If X is centrally symmetric distributed about some $z^* \in E$, then any depth function $D(\cdot \mid X)$ is maximal at z^*.*

Recall that X is *centrally symmetric* distributed about z^* if the distributions of $X - z^*$ and $z^* - X$ coincide.

Our definition of a depth function differs slightly from that given in Liu (1990) and Zuo and Serfling (2000). The main difference between these postulates and ours is that they additionally postulate Proposition 2.1 to be true and that they do not require upper semicontinuity **D5**.

D4 states that the upper level set $D_\alpha(x^1, \dots, x^n)$ are starshaped with respect to z^*. If a depth function, in place of **D4**, meets the restriction

- **D4con**: $D(\cdot \mid X)$ is a *quasiconcave* function, that is, its upper level sets $D_\alpha(X)$ are convex for all $\alpha > 0$,

the depth is mentioned as a *convex depth*. Obviously, as a convex set is starshaped with respect to each of its points, **D4con** implies **D4**. In certain settings the restriction **D2** is weakened to

- **D2iso**: $D(Az \mid AX) = D(z \mid X)$ for every *isometric linear* transformation $A : E \to E$.

Then, in case $E = \mathbb{R}^d$, D is called an *orthogonal invariant depth* in contrast to an *affine invariant* depth when **D2** holds. Alternatively, sometimes **D2** is attenuated to *scale invariance*,

- **D2sca**: $D(\lambda z \mid \lambda X) = D(z \mid X)$ for all $\lambda > 0$.

2.2.2 Central Regions and Outliers

For given P and $0 \le \alpha \le 1$, the level sets $D_\alpha(P)$ form a nested family of *depth-trimmed* or *central regions*. The innermost region arises at some $\alpha_{\max} \le 1$, which in general depends on P. $D_{\alpha_{\max}}(P)$ is the set of *deepest points*. **D1** and **D2** say that the family of central regions is affine equivariant. Central regions describe a distribution X with respect to location, dispersion, and shape. This has many applications in multivariate data analysis. On the other hand, given a nested family $\{C_\alpha(P)\}_{\alpha\in[0,1]}$ of set-valued statistics, defined on $\mathcal{P}$, that are convex, bounded and closed, the function D,

$$D\big(z \mid P\big) = \sup\big\{\alpha : z \in C_\alpha(P)\big\}, \quad z \in E,\ P \in \mathcal{P}, \tag{2.1}$$

satisfies **D1** to **D5** and **D4con**, hence is a convex depth function.

A depth function D orders data by their degree of centrality. Given a sample, it provides a center-outward *order statistic*. The depth induces an *outlyingness function* $\mathbb{R}^d \to [0, \infty[$ by

$$\mathrm{Out}\big(z \mid X\big) = \frac{1}{D(z \mid X)} - 1,$$

which is zero at the center and infinite at infinity. In turn, $D(z \mid X) = (1 + \mathrm{Out}(z \mid X))^{-1}$. Points outside a central region D_α have outlyingness greater than $1/\alpha - 1$; they can be regarded as *outliers* of a specified level α.

2.2.3 Depth Lifts, Stochastic Orderings, and Metrics

Assume $\alpha_{\max} = 1$ for $P \in \mathcal{P}$. By adding a real dimension to the central regions $D_\alpha(P), \alpha \in [0,1]$, we construct a set, which will be mentioned as the *depth lift*,

$$\hat{D}(P) = \big\{(\alpha, y) \in [0,1] \times E : y = \alpha x, x \in D_\alpha(P), \alpha \in [0,1]\big\}. \tag{2.2}$$

The depth lift gives rise to an *ordering* of probability distributions in $\mathcal{P}$: $P \prec_D Q$ if

$$\hat{D}(P) \subset \hat{D}(Q). \tag{2.3}$$

The restriction $\hat{D}(P) \subset \hat{D}(Q)$ is equivalent to $D_\alpha(P) \subset D_\alpha(Q)$ for all α. Thus, $P \prec_D Q$ means that each central set of Q is larger than the respective central set of P. In this sense, Q is *more dispersed* than P. The depth ordering is antisymmetric, hence an *order*, if and only if the family of central regions completely characterizes the underlying probability. Otherwise it is a preorder only. Finally, the depth D introduces a *probability semi-metric* on $\mathcal{P}$ by the Hausdorff distance of depth lifts,

$$\delta_D(P, Q) = \delta_H\big(\hat{D}(P), \hat{D}(Q)\big). \tag{2.4}$$

Recall that the *Hausdorff distance* $\delta_H(C_1, C_2)$ of two compact sets C_1 and C_2 is the smallest ε such that C_1 plus the ε-ball includes C_2 and vice versa. Again, the semi-metric is a metric iff the central regions characterize the probability.

2.3 Multivariate Depth Functions

Originally and in most existing applications depth statistics are used with data in Euclidean space. Multivariate depth statistics are particularly suited to analyze non-Gaussian or, more general, non-elliptical distributions in $\mathbb{R}^d$. Without loss of generality, we consider distributions of full dimension d, that is, whose convex hull of support, $\mathrm{co}(P)$, has affine dimension d.

A random vector X in $\mathbb{R}^d$ has a *spherical distribution* if AX is distributed as X for every orthogonal matrix A. It has an *elliptical distribution* if $X = a + BY$ for some $a \in \mathbb{R}^d$, $B \in \mathbb{R}^{d \times d}$, and spherically distributed Y; then we write $X \sim \mathrm{Ell}(a, BB', \varphi)$, where φ is the radial distribution of Y. Actually, on an elliptical distribution $P = \mathrm{Ell}(a, BB', \varphi)$, any depth function $D(\cdot, P)$ satisfying **D1** and **D2** has parallel elliptical level sets $D_\alpha(P)$, that is, level sets of a quadratic form with *scatter matrix* BB'. Consequently, all affine invariant depth functions are essentially equivalent if the distribution is elliptical. Moreover, if P is elliptical and has a unimodal Lebesgue-density f_P, the density level sets have the same elliptical shape, and the density is a transformation of the depth, i.e., a function φ exists such that $f_P(z) = \varphi(D(z \mid P))$ for all $z \in \mathbb{R}^d$. Similarly, on a spherical distribution, any depth satisfying postulates **D1** and **D2iso** has analogous properties.

In the following, we consider three principal approaches to define a multivariate depth statistic. The first approach is based on distances from properly defined central points or on volumes (Sect. 2.3.1), the second on certain L-statistics (*viz.* decreasingly weighted means of order statistics; Sect. 2.3.2), the third on simplices and halfspaces in $\mathbb{R}^d$ (Sect. 2.3.3). The three approaches have different consequences on the depths' ability to reflect asymmetries of the distribution, on their robustness to possible outliers, and on their computability with higher-dimensional data.

Figures 2.1, 2.2, 2.3 and 2.4 below exhibit bivariate central regions for several depths and equidistant α. The data consist of the unemployment rate (in %) and the GDP share of public debt for the countries of the European Union in 2011.

Most of the multivariate depths considered are convex and affine invariant, some exhibit spherical invariance only. Some are continuous in the point z or in the distribution P (regarding weak convergence), others are not. They differ in the shape of the depth lift and whether it uniquely determines the underlying distribution. A basic dispersion ordering of multivariate probability distributions serving as a benchmark is the *dilation order*, which says that Y spreads out more than X if $\mathrm{E}[\varphi(X)] \le \mathrm{E}[\varphi(Y)]$ holds for every convex $\varphi : \mathbb{R}^d \to \mathbb{R}$; see, e.g., Mosler (2002). It is interesting whether or not a particular depth ordering is concordant with the dilation order.

2.3.1 Depths Based on Distances

The outlyingness of a point, and hence its depth, can be measured by a distance from a properly chosen center of the distribution. In the following notions, this is done with different distances and centers.

***L_2*-Depth** The L_2-depth, D^{L_2}, is based on the mean outlyingness of a point, as measured by the L_2 distance,

$$D^{L_2}\big(z \mid X\big) = \big(1 + \mathrm{E}\|z - X\|\big)^{-1}. \tag{2.5}$$

It holds $\alpha_{\max} = 1$. The depth lift is $\hat{D}^{L_2}(X) = \{(\alpha, z) : \mathrm{E}\|z - \alpha X\| \le 1 - \alpha\}$ and convex. For an empirical distribution on points $x^i, i = 1, \dots, n$, we obtain

$$D^{L_2}\big(z \mid x^1, \dots, x^n\big) = \left(1 + \frac{1}{n}\sum_{i=1}^{n} \|z - x^i\|\right)^{-1}. \tag{2.6}$$

Obviously, the L_2-depth vanishes at infinity (**D3**), and is maximum at the *spatial median* of X, i.e., at the point $z \in \mathbb{R}^d$ that minimizes $\mathrm{E}\|z - X\|$. If the distribution is centrally symmetric, the center is the spatial median, hence the maximum is attained at the center. Monotonicity with respect to the deepest point (**D4**) as well as convexity and compactness of the central regions (**D4con**, **D5**) derive immediately from the triangle inequality. Further, the L_2-depth depends continuously on z. The L_2-depth converges also in the probability distribution: For a uniformly integrable and weakly convergent sequence $P_n \to P$ it holds $\lim_n D(z \mid P_n) = D(z \mid P)$.

However, the ordering induced by the L_2-depth is no sensible ordering of dispersion, since the L_2-depth contradicts the dilation order. As $\|z - x\|$ is convex in x, the expectation $\mathrm{E}\|z - X\|$ increases with a dilation of P. Hence, (2.5) decreases (!) with a dilation.

The L_2-depth is invariant against rigid Euclidean motions (**D1**, **D2iso**), but not affine invariant. An affine invariant version is constructed as follows: Given a positive definite $d \times d$ matrix M, consider the *M-norm*,

$$\|z\|_M = \sqrt{z' M^{-1} z}, \quad z \in \mathbb{R}^d. \tag{2.7}$$

Let S_X be a positive definite $d \times d$ matrix that depends continuously (in weak convergence) on the distribution and measures the dispersion of X in an affine equivariant way. The latter means that

$$S_{XA+b} = A S_X A' \text{ holds for any matrix } A \text{ of full rank and any } b. \tag{2.8}$$

Then an *affine invariant L_2-depth* is given by

$$\big(1 + \mathrm{E}\|z - X\|_{S_X}\big)^{-1}. \tag{2.9}$$

Besides invariance, it has the same properties as the L_2-depth. A simple choice for S_X is the covariance matrix Σ_X of X (Zuo and Serfling 2000). Note that the covariance matrix is positive definite, as the convex hull of the support, co(P), is assumed to have full dimension. More robust choices for S_X are the *minimum volume ellipsoid* (MVE) or the *minimum covariance determinant* (MCD) estimators; see Rousseeuw and Leroy (1987), Lopuhaä and Rousseeuw (1991), and the contribution by Rousseeuw and Hubert, Chap. 4.

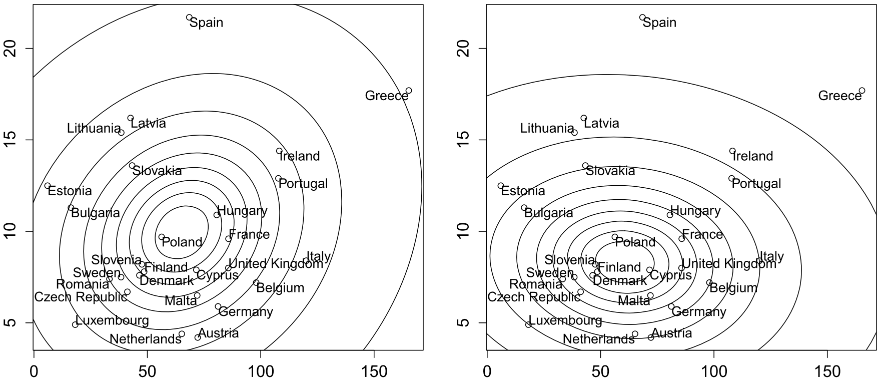


Fig. 2.1 Governmental debt (x-axis) and unemployment rate (y-axis); Mahalanobis regions (moment, *left*; MCD, *right*) with $\alpha = 0.1(0.1), \dots, 0.9$

Mahalanobis Depths Let c_X be a vector that measures the location of X in a continuous and affine equivariant way and, as before, S_X be a matrix that satisfies (2.8) and depends continuously on the distribution. Based on the estimates c_X and S_X a simple depth statistic is constructed, the *generalized Mahalanobis depth*, given by

$$D^{\text{Mah}}\big(z \mid X\big) = \big(1 + \|z - c_X\|^2_{S_X}\big)^{-1}. \tag{2.10}$$

Obviously, (2.10) satisfies **D1** to **D5** and **D4con**, taking its unique maximum at c_X. The depth lift is the convex set $\hat{D}^{\text{Mah}}(X) = \{(\alpha, z) : \|z - \alpha c_X\|^2_{S_X} \leq \alpha^2(\alpha - 1)\}$, and the central regions are ellipsoids around c_X. The generalized Mahalanobis depth is continuous on z and P. In particular, with $c_X = \mathrm{E}[X]$ and $S_X = \Sigma_X$ the *(moment) Mahalanobis depth* is obtained,

$$D^{\text{mMah}}\big(z \mid X\big) = \big(1 + \big(z - \mathrm{E}[X]\big)' \Sigma_X^{-1} \big(z - \mathrm{E}[X]\big)\big)^{-1}. \tag{2.11}$$

Its sample version is

$$D^{\text{mMah}}\big(z \mid x^1, \dots, x^n\big) = \big(1 + (z - \bar{x})' \hat{\Sigma}_x^{-1} (z - \bar{x})\big)^{-1}, \tag{2.12}$$

where $\bar{x}$ is the mean vector and $\hat{\Sigma}_X$ is the empirical covariance matrix. It is easily seen that the α-central set of a sample from P converges almost surely to the α-central set of P, for any α. Figure 2.1 shows Mahalanobis regions for the debt-unemployment data, employing two choices of the matrix S_X, namely the usual moment estimate Σ_X and the robust MCD estimate. As it is seen from the Figure, these regions depend heavily on the choice of S_X. Hungary, e.g., is rather central (having depth greater than 0.8) with the moment Mahalanobis depth, while it is much more outlying (having depth below 0.5) with the MCD version.

Concerning uniqueness, the Mahalanobis depth fails in identifying the underlying distribution. As only the first two moments are used, any two distributions which have the same first two moments cannot be distinguished by their Mahalanobis depth functions. Similarly, the generalized Mahalanobis depth does not determine the distribution. However, within the family of nondegenerate d-variate normal distributions or, more general, within any affine family of nondegenerate d-variate distributions having finite second moments, a single contour set of the Mahalanobis depth suffices to identify the distribution.

Projection Depth The *projection depth* has been proposed in Zuo and Serfling (2000):

$$D^{\text{proj}}\big(z \mid X\big) = \left(1 + \sup_{p \in S^{d-1}} \frac{|\langle p, z\rangle - \text{med}(\langle p, X\rangle)|}{\text{Dmed}(\langle p, X\rangle)}\right)^{-1}, \tag{2.13}$$

where S^{d-1} denotes the unit sphere in $\mathbb{R}^d$, $\langle p, z\rangle$ is the inner product (that is the projection of z on the line $\{\lambda p : \lambda \in \mathbb{R}\}$), $\text{med}(U)$ is the usual median of a univariate random variable U, and $\text{Dmed}(U) = \text{med}(|U - \text{med}(U)|)$ is the median absolute deviation from the median. The projection depth satisfies **D1** to **D5** and **D4con**. It has good properties, which are discussed in detail by Zuo and Serfling (2000). For breakdown properties of the employed location and scatter statistics, see Zuo (2000).

Oja Depth The Oja depth is not based on distances, but on average volumes of simplices that have vertices from the data (Zuo and Serfling 2000):

$$D^{\text{Oja}}\big(z \mid X\big) = \left(1 + \frac{\text{E}(\text{vol}_d(\text{co}\{z, X_1, \ldots, X_d\}))}{\sqrt{\det \Sigma_X}}\right)^{-1},$$

where $X_1, \ldots, X_d$ are random vectors independently distributed as P, co denotes the convex hull, V_d the d-dimensional volume, and S_X is defined as above. In particular, we can choose $D_X = \Sigma_X$. The Oja depth satisfies **D1** to **D5**. It is continuous on z and maximum at the Oja median (Oja 1983), which is not unique; see also the contribution by Oja, Chap. 1. The Oja depth determines the distribution uniquely among those measures which have compact support of full dimension.

Figure 2.2 contrasts the projection depth regions with the Oja regions for our debt-unemployment data. The regions have different shapes, but agree in making Spain and Greece the most outlying countries.

2.3.2 Weighted Mean Depths

A large and flexible class of depth statistics corresponds to so called weighted-mean central regions, shortly WM regions (Dyckerhoff and Mosler 2011, 2012). These are convex compacts in $\mathbb{R}^d$, whose support function is a weighted mean of order statistics, that is, an L-statistic. Recall that a convex compact $K \subset \mathbb{R}^d$ is uniquely determined by its support function h_K,

$$h_K(p) = \max\big\{p'x : x \in K\big\}, \quad p \in S^{d-1}.$$

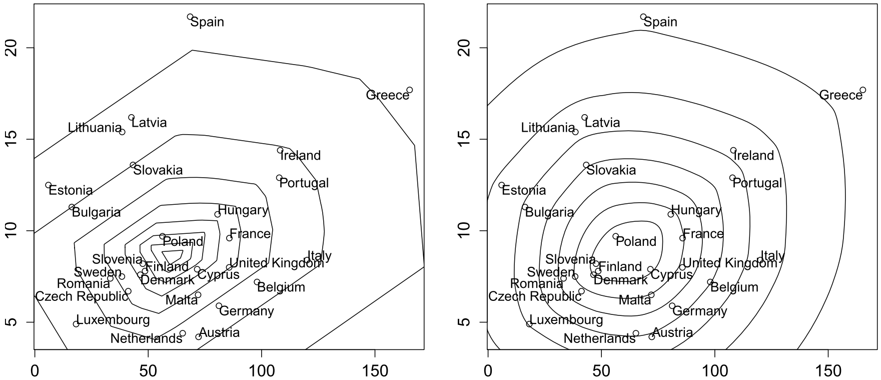


Fig. 2.2 Governmental debt and unemployment rate; projection depth regions (*left*), Oja regions (*right*); both with $\alpha = 0.1(0.1), \dots, 0.9$

To define the WM α-region of an empirical distribution on $x^1, x^2, \dots, x^n$, we construct its support function as follows: For $p \in S^{d-1}$, consider the line $\{\lambda p \in \mathbb{R}^d : \lambda \in \mathbb{R}\}$. By projecting the data on this line a linear ordering is obtained,

$$p'x^{\pi_p(1)} \le p'x^{\pi_p(2)} \le \dots \le p'x^{\pi_p(n)}, \tag{2.14}$$

and, by this, a permutation π_p of the indices $1, 2, \dots, n$. Consider weights $w_{j,\alpha}$ for $j \in \{1, 2, \dots, n\}$ and $\alpha \in [0, 1]$ that satisfy the following restrictions (i) to (iii):

(i) $\sum_{j=1}^n w_{j,\alpha} = 1$, $w_{j,\alpha} \ge 0$ for all j and α.
(ii) $w_{j,\alpha}$ increases in j for all α.
(iii) $\alpha < \beta$ implies $\sum_{j=1}^k w_{j,\alpha} \le \sum_{j=1}^k w_{j,\beta}$, $k = 1, \dots, n$.

Then, as it has been shown in Dyckerhoff and Mosler (2011), the function $h_{D_\alpha(x^1,\dots,x^n)}$,

$$h_{D_\alpha(x^1,\dots,x^n)}(p) = \sum_{j=1}^n w_{j,\alpha} p' x^{\pi_p(j)}, \quad p \in S^{d-1}, \tag{2.15}$$

is the support function of a convex body $D_\alpha = D_\alpha(x^1, \dots, x^n)$, and $D_\alpha \subset D_\beta$ holds whenever $\alpha > \beta$. Now we are ready to see the general definition of a family of WM regions.

Definition 2.1 Given a weight vector $w_\alpha = w_{1,\alpha}, \dots, w_{n,\alpha}$ that satisfies the restrictions (i) to (iii), the convex compact $D_\alpha = D_\alpha(x^1, \dots, x^n)$ having support function (2.15) is named the *WM region* of $x^1, \dots, x^n$ at level α, $\alpha \in [0, 1]$. The corresponding depth (2.1) is the *WM depth* with weights w_α, $\alpha \in [0, 1]$.

It follows that the WM depth satisfies the restrictions **D1** to **D5** and **D4con**. Moreover, it holds

$$D_\alpha\bigl(x^1,\dots,x^n\bigr) = \operatorname{conv}\left\{\sum_{j=1}^n w_{j,\alpha} x^{\pi(j)} : \pi \text{ permutation of } \{1,\dots,n\}\right\}. \tag{2.16}$$

This explains the name by stating that a WM region is the convex hull of weighted means of the data. Consequently, outside the convex hull of the data the WM depth vanishes. WM depths are useful statistical tools as their central regions have attractive analytical and computational properties. Sample WM regions are consistent estimators for the WM region of the underlying probability. Besides being *continuous* in the distribution and in α, WM regions are *subadditive*, that is,

$$D_\alpha\bigl(x^1+y^1,\dots,x^n+y^n\bigr) \subset D_\alpha\bigl(x^1,\dots,x^n\bigr) \oplus D_\alpha\bigl(y^1,\dots,y^n\bigr),$$

and *monotone*: If $x^i \le y^i$ holds for all i (in the componentwise ordering of $\mathbb{R}^d$), then

$$\begin{aligned} &D_\alpha\bigl(y^1,\dots,y^n\bigr) \subset D_\alpha\bigl(x^1,\dots,x^n\bigr) \oplus \mathbb{R}^d_+ \quad \text{and} \\ &D_\alpha\bigl(x^1,\dots,x^n\bigr) \subset D_\alpha\bigl(y^1,\dots,y^n\bigr) \oplus \mathbb{R}^d_-, \end{aligned}$$

where $\oplus$ signifies the Minkowski sum of sets.

Depending on the choice of the weights $w_{j,\alpha}$ different notions of data depths are obtained. For a detailed discussion of these and other special WM depths and central regions, the reader is referred to Dyckerhoff and Mosler (2011, 2012).

Zonoid Depth For an empirical distribution P on $x^1,\dots,x^n$ and $0<\alpha\le 1$ define the zonoid region (Koshevoy and Mosler 1997)

$$D_\alpha^{\text{zon}}(P) = \left\{\sum_{i=1}^n \lambda_i x^i : 0 \le \lambda_i \le \frac{1}{n\alpha},\ \sum_{i=1}^n \lambda_i = 1\right\}.$$

See Fig. 2.3. The corresponding support function (2.15) employs the weights

$$w_{j,\alpha} = \begin{cases} 0 & \text{if } j < n - \lfloor n\alpha \rfloor, \\ \frac{n\alpha - \lfloor n\alpha \rfloor}{n\alpha} & \text{if } j = n - \lfloor n\alpha \rfloor, \\ \frac{1}{n\alpha} & \text{if } j > n - \lfloor n\alpha \rfloor. \end{cases} \tag{2.17}$$

Many properties of zonoid regions and the zonoid depth $D^{\text{zon}}(z \mid X)$ are discussed in Mosler (2002). The zonoid depth lift equals the so called lift zonoid, which fully characterizes the distribution. Therefore the zonoid depth generates an antisymmetric depth order (2.3) and a probability metric (2.4). Zonoid regions are not only invariant to affine, but to general linear transformations; specifically any marginal projection of a zonoid region is the zonoid region of the marginal distribution. The zonoid depth is continuous on z as well as P.

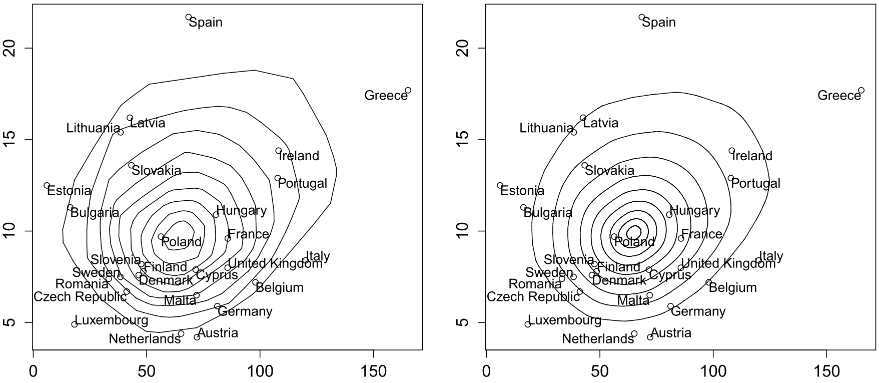


Fig. 2.3 Governmental debt and unemployment rate; zonoid regions (*left*), ECH* regions (*right*); both with $\alpha = 0.1(0.1), \ldots, 0.9$

Expected Convex Hull Depth Another important notion of WMT depth is that of *expected convex hull (ECH*)* depth (Cascos 2007). Its central region D_α (see Fig. 2.3) has a support function with weights

$$w_{j,\alpha} = \frac{j^{1/\alpha} - (j-1)^{1/\alpha}}{n^{1/\alpha}}. \tag{2.18}$$

Figure 2.3 depicts zonoid and ECH* regions for our data. We see that the zonoid regions are somewhat angular while the ECH* regions appear to be smoother; this corresponds, when calculating such regions in higher dimensions, to a considerably higher computation load of ECH*.

Geometrical Depth The weights

$$w_{j,\alpha} = \begin{cases} \frac{1-\alpha}{1-\alpha^n}\alpha^{n-j} & \text{if } 0 < \alpha < 1, \\ 0 & \text{if } \alpha = 1, \end{cases}$$

yield another class of WM regions. The respective depth is the *geometrically weighted mean depth* (Dyckerhoff and Mosler 2011).

2.3.3 Depths Based on Halfspaces and Simplices

The third approach concerns no distances or volumes, but the combinatorics of halfspaces and simplices only. In this it is independent of the metric structure of $\mathbb{R}^d$. While depths that are based on distances or weighted means may be addressed as *metric depths*, the following ones will be mentioned as *combinatorial depths*. They remain constant, as long as the compartment structure of the data does not change. By this, they are very robust against *location outliers*. Outside the convex support $\text{co}(X)$ of the distribution every combinatorial depth attains its minimal value, which is zero.

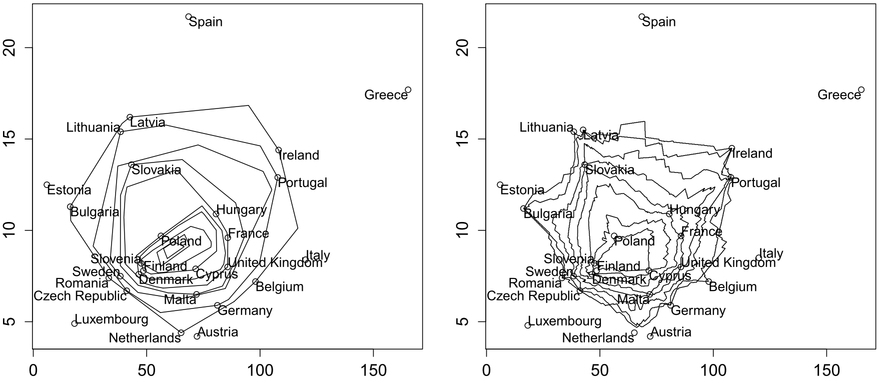


Fig. 2.4 Governmental debt and unemployment rate; Tukey regions (*left*) with $\alpha = \frac{2}{27}(\frac{1}{27})$, $\dots, \frac{1}{1}27$, simplicial regions (*right*) with $\alpha = 0.25, 0.3(0.1), \dots, 0.9$

Location Depth Consider the population version of the *location depth*,

$$D^{\mathrm{loc}}\big(z \mid X\big) = \inf\big\{P(H) : H \text{ is a closed halfspace, } z \in H\big\}. \tag{2.19}$$

The depth is also known as *halfspace* or *Tukey depth*, its central regions as *Tukey regions*. The location depth is affine invariant (**D1**, **D2**). Its central regions are convex (**D4con**) and closed (**D5**); see Fig. 2.4. The maximum value of the location depth is smaller or equal to 1 depending on the distribution. The set of all such points is mentioned as the *halfspace median set* and each of its elements as a *Tukey median* (Tukey 1975).

If X has an *angular symmetric* distribution, the location depth attains its maximum at the center and the center is a Tukey median; this strengthens Proposition 2.1. (A distribution is called *angular* (= *halfspace*) *symmetric* about z^* if $P(X \in H) \geq 1/2$ for every closed halfspace H having z^* on the boundary; equivalently, if $(X - z^*)/\|X - z^*\|$ is centrally symmetric with the convention $0/0 = 0$.)

If X has a Lebesgue-density, the location depth depends continuously on z; otherwise the dependence on z is noncontinuous and there can be more than one point where the maximum is attained. As a function of P the location depth is obviously noncontinuous. It determines the distribution in a unique way if the distribution is either discrete (Struyf and Rousseeuw 1999; Koshevoy 2002) or continuous with compact support. The location depth of a sample from P converges almost surely to the location depth of P (Donoho and Gasko 1992). The next depth notion involves simplices in $\mathbb{R}^d$.

Simplicial Depth Liu (1990) defines the *simplicial depth* as follows:

$$D^{\mathrm{sim}}\big(z \mid X\big) = P\big(z \in \mathrm{co}\big(\{X_1, \dots, X_{d+1}\}\big)\big), \tag{2.20}$$

where $X_1, \dots, X_{d+1}$ are i.i.d. by P. The sample version reads as

$$D^{\mathrm{sim}}\big(z \mid x^1, \dots, x^n\big) = \frac{1}{\binom{n}{d+1}}\, \#\big\{\{i_1, \dots, i_{d+1}\} : z \in \mathrm{co}\big(\big\{x^{i_1}, \dots, x^{i_{d+1}}\big\}\big)\big\}. \tag{2.21}$$

The simplicial depth is affine invariant (**D1**, **D2**). Its maximum is less or equal to 1, depending on the distribution. In general, the point of maximum simplicial depth is not unique; the *simplicial median* is defined as the gravity center of these points. The sample simplicial depth converges almost surely uniformly in z to its population version (Liu 1990; Dümbgen 1992). The simplicial depth has positive breakdown (Chen 1995).

If the distribution is Lebesgue-continuous, the simplicial depth behaves well: It varies continuously on z (Liu 1990, Theorem 2), is maximum at a center of angular symmetry, and decreases monotonously from a deepest point (**D4**). The *simplicial central regions* of a Lebesgue-continuous distribution are connected and compact (Liu 1990).

However, if the distribution is discrete, each of these properties can fail; for counterexamples see, e.g., Zuo and Serfling (2000). The simplicial depth characterizes an empirical measure if the supporting points are in *general position*, that is, if no more than d of the points lie on the same hyperplane.

As Fig. 2.4 demonstrates, Tukey regions are convex while simplicial regions are only starshaped. The figure illustrates also that these notions are rather insensitive to outlying data: both do not reflect *how far* Greece and Spain are from the center. Whether, in an application, this kind of robustness is an advantage or not, depends on the problem and data at hand.

Other well known combinatorial data depths are the *majority depth* (Liu and Singh 1993) and the *convex-hull peeling depth* (Barnett 1976; Donoho and Gasko 1992). However, the latter possesses no population version.

2.4 Functional Data Depth

The analysis of functional data has become a practically important branch of statistics; see Ramsay and Silverman (2005). Consider a space E of functions $[0,1] \to \mathbb{R}$ with the supremum norm. Like a multivariate data depth, a functional data depth is a real-valued functional that indicates how 'deep' a function $z \in E$ is located in a given finite cloud of functions $\in E$. Let E' denote the set of continuous linear functionals $E \to \mathbb{R}$, and E'^d the d-fold Cartesian product of E'. Here, following Mosler and Polyakova (2012), functional depths of a general form (2.22) are presented. Some alternative approaches will be addressed below.

***Φ*-Depth** For $z \in E$ and an empirical distribution X on $x^1, \dots, x^n \in E$, define a *functional data depth* by

$$D\big(z \mid X\big) = \inf_{\varphi \in \Phi} D^d\big(\varphi(z) \mid \varphi(X)\big), \tag{2.22}$$

where D^d is a d-variate data depth satisfying **D1** to **D5**, $\Phi \subset E'^d$, and $\varphi(X)$ is the empirical distribution on $\varphi(x^1), \dots, \varphi(x^n)$. D is called a *Φ-depth*. A population version is similarly defined.

Each φ in this definition may be regarded as a particular 'aspect' we are interested in and which is represented in d-dimensional space. The depth of z is given as the smallest multivariate depth of z under all these aspects. It implies that all aspects are equally relevant so that the depth of z cannot be larger than its depth under any aspect.

As the d-variate depth D^d has maximum not greater than 1, the functional data depth D is bounded above by 1. At every point z^* of *maximal D-depth* it holds $D(z^* \mid X) \le 1$. The bound is attained with equality, $D(z^* \mid X) = 1$, iff $D^d(\varphi(z^*) \mid \varphi(X)) = 1$ holds for all $\varphi \in \Phi$, that is, iff

$$z^* \in \bigcap_{\varphi\in\Phi} \varphi^{-1}\big(D_1^d\big(\varphi(X)\big)\big). \tag{2.23}$$

A Φ-depth (2.22) always satisfies **D1**, **D2sca**, **D4**, and **D5**.

It satisfies **D3** if for every sequence (z^i) with $\|z^i\| \to \infty$ exists a φ in Φ such that $\varphi(z^i) \to \infty$. (For some special notions of functional data depth this postulate has to be properly adapted.)

D4con is met if **D4con** holds for the underlying d-variate depth.

We now proceed with specifying the set Φ of functionals and the multivariate depth D^k in (2.22). While many features of the functional data depth (2.22) resemble those of a multivariate depth, an important difference must be pointed out: In a general Banach space the unit ball B is not compact, and properties **D3** and **D5** do not imply that the level sets of a functional data depth are compact. So, to obtain a meaningful notion of functional data depth of type (2.22) one has to carefully choose a set of functions Φ which is not too large. On the other hand, Φ should not be too small, in order to extract sufficient information from the data.

Graph Depths For $x \in E$ denote $x(t) = (x_1(t), \dots, x_d(t))$ and consider

$$\Phi = \left\{\varphi^t : E \to \mathbb{R}^d : \varphi^t(x) = \big(x_1(t), \dots, x_d(t)\big), t \in T\right\} \tag{2.24}$$

for some $T \subset [0, 1]$, which may be a subinterval or a finite set. For D^d use any multivariate depth that satisfies **D1** to **D5**. This results in the *graph depth*

$$\mathrm{GD}\big(z \mid x^1, \dots, x^n\big) = \inf_{t\in T} D^d\big(z(t) \mid x^1(t), \dots, x^n(t)\big). \tag{2.25}$$

In particular, with the univariate halfspace depth, $d = 1$ and $T = J$ we obtain the *halfgraph depth* (López-Pintado and Romo 2005). Also, with the univariate simplicial depth the *band depth* (López-Pintado and Romo 2009) is obtained, but this, in general, violates monotonicity **D4**.

Grid Depths We choose a finite number of points in J, $t_1, \dots, t_k$, and evaluate a function $z \in E$ at these points. Notate $\underline{t} = (t_1, \dots, t_k)$ and $z(\underline{t}) = (z_1(\underline{t}), \dots, z_d(\underline{t}))^{\mathsf{T}}$. That is, in place of the function z the $k \times d$ matrix $z^{(k)}$ is considered. A *grid depth* RD is defined by (2.22) with the following Φ,

$$\Phi = \left\{\varphi^r : \varphi^r(z) = \big(\langle r, z_1(\underline{t})\rangle, \dots, \langle r, z_d(\underline{t})\rangle\big), r \in S^{k-1}\right\}, \tag{2.26}$$

which yields

$$\mathrm{RD}\big(z \mid x^1, \dots, x^n\big) = \inf_{r \in S^{k-1}} D^d\big(\langle r, z(\underline{t})\rangle \mid \langle r, x^1(\underline{t})\rangle, \dots, \langle r, x^n(\underline{t})\rangle\big). \tag{2.27}$$

A slight extension of the Φ-depth is the *principal components depth* (Mosler and Polyakova 2012). However, certain approaches from the literature are no Φ-depths. These are mainly of two types. The first type employs *random projections* of the data: Cuesta-Albertos and Nieto-Reyes (2008b) define the depth of a function as the univariate depth of the function values taken at a randomly chosen argument t. Cuevas et al. (2007) also employ a random projection method. The other type uses average univariate depths. Fraiman and Muniz (2001) calculate the univariate depths of the values of a function and integrate them over the whole interval; this results in kind of 'average' depth. Claeskens et al. (2012) introduce a multivariate ($d \geq 1$) functional data depth, where they similarly compute a weighted average depth. The weight at a point reflects the variability of the function values at this point (more precisely: is proportional to the volume of a central region at the point).

2.5 Computation of Depths and Central Regions

The moment Mahalanobis depth and its elliptical central regions are obtained in any dimension by calculating the mean and the sample covariance matrix, while robust Mahalanobis depths and regions are determined with the R-procedures "cov.mcd" and "cov.mve". In dimension $d = 2$, the central regions of many depth notions can be exactly calculated by following a *circular sequence* (Edelsbrunner 1987). The R-package "depth" computes the exact location ($d = 2, 3$) and simplicial ($d = 2$) depths, as well as the Oja depth and an approximative location depth for any dimension. An exact algorithm for the location depth in any dimension is developed in Liu and Zuo (2012). Cuesta-Albertos and Nieto-Reyes (2008a) propose to calculate instead the *random Tukey depth*, which is the minimum univariate location depth of univariate projections in a number of randomly chosen directions. With the algorithm of Paindaveine and Šiman (2012), Tukey regions are obtained, $d \geq 2$. The bivariate projection depth is computed by the R-package "ExPD2D"; for the respective regions, see Liu et al. (2011). The zonoid depth can be efficiently determined in any dimension (Dyckerhoff et al. 1996). An R-package ("WMTregions") exists for the exact calculation of zonoid and general WM regions; see Mosler et al. (2009), Bazovkin and Mosler (2012). The R-package "rainbow" calculates several functional data depths.

2.6 Conclusions

Depth statistics have been used in numerous and diverse tasks of which we can mention a few only. Liu et al. (1999) provide an introduction to some of them. In descriptive multivariate analysis, depth functions and central regions visualize the data regarding location, scale and shape. By bagplots and sunburst plots, outliers can be identified and treated in an interactive way. In k-class supervised classification, each—possibly high-dimensional—data point is represented in $[0, 1]^k$ by its values of depth in the k given classes, and classification is done in $[0, 1]^k$. Functions of depth statistics include depth-weighted statistical functionals, such as $\int_{\mathbb{R}^d} x w(D(x \mid P))\, dP / \int_{\mathbb{R}^d} w(D(x \mid P))\, dP$ for location. In inference, tests for goodness of fit and homogeneity regarding location, scale and symmetry are based on depth statistics; see, e.g., Dyckerhoff (2002), Ley and Paindaveine (2011). Applications include such diverse fields as statistical control (Liu and Singh 1993), measurement of risk (Cascos and Molchanov 2007), and robust linear programming (Bazovkin and Mosler 2011). Functional data depth is applied to similar tasks in description, classification and testing; see, e.g., López-Pintado and Romo (2009), Cuevas et al. (2007).

This survey has covered the fundamentals of depth statistics for d-variate and functional data. Several special depth functions in $\mathbb{R}^d$ have been presented, metric and combinatorial ones, with a focus on the recent class of WM depths. For functional data, depths of infimum type have been discussed. Of course, such a survey is necessarily incomplete and biased by the preferences of the author. Of the many applications of depth in the literature only a few have been touched, and important theoretical extensions like regression depth (Rousseeuw and Hubert 1999), depth calculus (Mizera 2002), location-scale depth (Mizera and Müller 2004), and likelihood depth (Müller 2005) have been completely omitted.

Most important for the selection of a depth statistic in applications are the questions of computability and—depending on the data situation—robustness. Mahalanobis depth is solely based on estimates of the mean vector and the covariance matrix. In its classical form with moment estimates Mahalanobis depth is efficiently calculated but highly non-robust, while with estimates like the minimum volume ellipsoid it becomes more robust. However, since it is constant on ellipsoids around the center, Mahalanobis depth cannot reflect possible asymmetries of the data. Zonoid depth can be efficiently calculated, also in larger dimensions, but has the drawback that the deepest point is always the mean, which makes the depth non-robust. So, if robustness is an issue, the zonoid depth has to be combined with a proper preprocessing of the data to identify possible outliers. The location depth is, by construction, very robust but expensive when exactly computed in dimensions more than two. As an efficient approach the random Tukey depth yields an upper bound on the location depth, where the number of directions has to be somehow chosen.

A depth statistics measures the centrality of a point in the data. Besides ordering the data it provides numerical values that, with some depth notions, have an obvious meaning; so with the location depth and all WM depths. With other depths, in particular those based on distances, the outlyingness function has a direct interpretation.

References

Barnett, V. (1976). The ordering of multivariate data (with discussion). *Journal of the Royal Statistical Society. Series A. General*, *139*, 318–352.

Bazovkin, P., & Mosler, K. (2011). *Stochastic linear programming with a distortion risk constraint*. arXiv:1208.2113v1.

Bazovkin, P., & Mosler, K. (2012). An exact algorithm for weighted-mean trimmed regions in any dimension. *Journal of Statistical Software*, *47*(13).

Cascos, I. (2007). The expected convex hull trimmed regions of a sample. *Computational Statistics*, *22*, 557–569.

Cascos, I. (2009). Data depth: multivariate statistics and geometry. In W. Kendall & I. Molchanov (Eds.), *New perspectives in stochastic geometry*, Oxford: Clarendon/Oxford University Press.

Cascos, I., & Molchanov, I. (2007). Multivariate risks and depth-trimmed regions. *Finance and Stochastics*, *11*, 373–397.

Chen, Z. (1995). Bounds for the breakdown point of the simplicial median. *Journal of Multivariate Analysis*, *55*, 1–13.

Claeskens, G., Hubert, M., & Slaets, L. (2012). Multivariate functional halfspace depth. In *Workshop robust methods for dependent data*, Witten.

Cuesta-Albertos, J., & Nieto-Reyes, A. (2008a). A random functional depth. In S. Dabo-Niang & F. Ferraty (Eds.), *Functional and operatorial statistics* (pp. 121–126). Heidelberg: Physica-Verlag.

Cuesta-Albertos, J., & Nieto-Reyes, A. (2008b). The random Tukey depth. *Computational Statistics & Data Analysis*, *52*, 4979–4988.

Cuevas, A., Febrero, M., & Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, *22*, 481–496.

Donoho, D. L., & Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, *20*, 1803–1827.

Dümbgen, L. (1992). Limit theorems for the simplicial depth. *Statistics & Probability Letters*, *14*, 119–128.

Dyckerhoff, R. (2002). *Datentiefe: Begriff, Berechnung, Tests*. Mimeo, Fakultät für Wirtschafts- und Sozialwissenschaften, Universität zu Köln.

Dyckerhoff, R., Koshevoy, G., & Mosler, K. (1996). Zonoid data depth: theory and computation. In A. Pratt (Ed.), *Proceedings in computational statistics COMPSTAT* (pp. 235–240). Heidelberg: Physica-Verlag.

Dyckerhoff, R., & Mosler, K. (2011). Weighted-mean trimming of multivariate data. *Journal of Multivariate Analysis*, *102*, 405–421.

Dyckerhoff, R., & Mosler, K. (2012). Weighted-mean regions of a probability distribution. *Statistics & Probability Letters*, *82*, 318–325.

Edelsbrunner, H. (1987). *Algorithms in combinatorial geometry*. Heidelberg: Springer.

Fraiman, R., & Muniz, G. (2001). Trimmed means for functional data. *Test*, *10*, 419–440.

Koshevoy, G. (2002). The Tukey depth characterizes the atomic measure. *Journal of Multivariate Analysis*, *83*, 360–364.

Koshevoy, G., & Mosler, K. (1997). Zonoid trimming for multivariate distributions. *The Annals of Statistics*, *25*, 1998–2017.

Ley, C., & Paindaveine, D. (2011). *Depth-based runs tests for multivariate central symmetry*. ECARES discussion papers 2011/06, ULB, Bruxelles.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, *18*, 405–414.

Liu, R. Y., Parelius, J. M., & Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussion). *The Annals of Statistics*, *27*, 783–858.

Liu, R. Y., Serfling, R., & Souvaine, D. L. (2006). *Data depth: robust multivariate analysis, computational geometry and applications*. Providence: Am. Math. Soc.

Liu, R. Y., & Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, *88*, 252–260.

Liu, X., & Zuo, Y. (2012). *Computing halfspace depth and regression depth*. Mimeo.

Liu, X., Zuo, Y., & Wang, Z. (2011). *Exactly computing bivariate projection depth contours and median*. arXiv:1112.6162v1.

López-Pintado, S., & Romo, J. (2005). *A half-graph depth for functional data*. Working papers 01/2005, Universidad Carlos III, Statistics and Econometrics.

López-Pintado, S., & Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, *104*, 718–734.

Lopuhaä, H. P., & Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, *19*, 229–248.

Mizera, I. (2002). On depth and deep points: a calculus. *The Annals of Statistics*, *30*, 1681–1736.

Mizera, I., & Müller, C. H. (2004). Location-scale depth. *Journal of the American Statistical Association*, *99*, 949–989.

Mosler, K. (2002). *Multivariate dispersion, central regions and depth: the lift zonoid approach*. New York: Springer.

Mosler, K., Lange, T., & Bazovkin, P. (2009). Computing zonoid trimmed regions in dimension $d > 2$. *Computational Statistics & Data Analysis*, *53*, 2500–2510.

Mosler, K., & Polyakova, Y. (2012). *General notions of depth for functional data*. arXiv:1208.1981v1.

Müller, C. H. (2005). Depth estimators and tests based on the likelihood principle with application to regression. *Journal of Multivariate Analysis*, *95*, 153–181.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, *1*, 327–332.

Paindaveine, D., & Šiman, M. (2012). Computing multiple-output regression quantile regions. *Computational Statistics & Data Analysis*, *56*, 840–853.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). New York: Springer.

Rousseeuw, P. J., & Hubert, M. (1999). Regression depth (with discussion). *Journal of the American Statistical Association*, *94*, 388–433.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Serfling, R. (2006). Depth functions in nonparametric multivariate inference. In R. Liu, R. Serfling, & D. Souvaine (Eds.), *Data depth: robust multivariate analysis, computational geometry and applications* (pp. 1–16). Providence: Am. Math. Soc.

Struyf, A., & Rousseeuw, P. J. (1999). Halfspace depth and regression depth characterize the empirical distribution. *Journal of Multivariate Analysis*, *69*, 135–153.

Tukey, J. W. (1975). Mathematics and picturing data. In R. James (Ed.), *Proceedings of the 1974 international congress of mathematicians*, Vancouver (Vol. 2, pp. 523–531).

Zuo, Y. (2000). A note on finite sample breakdown points of projection based multivariate location and scatter statistics. *Metrika*, *51*, 259–265.

Zuo, Y., & Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, *28*, 461–482.

Chapter 3
Multivariate Extremes: A Conditional Quantile Approach

Marie-Françoise Barme-Delcroix

3.1 Introduction

Ordering multivariate data can be done in various ways and many definitions have been proposed by, e.g., Barnett (1976), Oja (1983), Maller (1990), Heffernan and Tawn (2004), Falk and Reiss (2005); see also the contribution by Oja, Chap. 1. Some papers of Einmahl and Mason (1992), Abdous and Theodorescu (1992), De Haan and Huang (1995), Berlinet et al. (2001), Serfling (2002), and more recently Hallin et al. (2010) develop the notion of multivariate quantiles. In the classical scheme (cartesian coordinates), the multivariate variables are ordered coordinate by coordinate—see for example Galambos (1987) and the references therein. And in this way the maximum value thus obtained is not a sample point. A new notion for the order statistics of a multivariate sample has been explored in Delcroix and Jacob (1991) by using the isobar-surfaces, that is, the level surfaces of the conditional distribution function of the radius given the angle. The sample is ordered relatively to an increasing family of isobars and the maximum value of the sample is the point of the sample belonging to the upper level isobar. This approach is more geometric and the maximum value is a sample point. The definition depends only on the conditional radial distribution. The first motivation was to describe the overall shape of a multidimensional sample, Barme-Delcroix (1993), and has given a new interest to the notion of stability, Geffroy (1958, 1961). By a unidimensional approach, some results have been stated in this multidimensional context such almost sure stability and strong behaviour, Barme-Delcroix and Brito (2001), or limit laws, Barme-Delcroix and Gather (2007). In Ivanková (2010), isobars are estimated by non-parametric regression methods and used to evaluate the efficiency of selected markets based on returns of their stock market indices.

This contribution is concerned with the theory of isobars. First, in the next section we recall some definitions and notations which will be useful throughout this paper.

M.-F. Barme-Delcroix (✉)
Laboratoire Painlevé, Lille1 University, Lille, France
e-mail: marie-francoise.barme@math.univ-lille1.fr

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_3, © Springer-Verlag Berlin Heidelberg 2013

In Sect. 3.3, as an introduction to the isobar-surfaces ordering, we give some results about the weak stability of this kind of multivariate extremes. This notion appears, as in the unidimensional case, strongly related to the notion of outlier-proneness or outlier resistance, Barme-Delcroix and Gather (2002). In Sect. 3.4, we propose a definition for the record times and record values of a multidimensional sequence of random variables, based on this isobar-surfaces ordering. At last in Sect. 3.5, we provide definitions of the stability for record values of multidimensional sequences and study the resulting probabilistic properties. The idea behind the definition is to describe the tendency of the record values to be near a given surface. We provide then characterizations, in term of the distribution function, for stability properties of the record values, as available in the univariate case, Resnick (1973a,b).

3.2 Preliminaries

Let X be an $\mathbf{R}^d$-valued random variable defined on a probability space $(\Omega, \mathcal{A}, P)$. Denote by $\|\cdot\|$ the Euclidean norm of $\mathbf{R}^d$ and by $\mathbf{S}^{d-1}$ the unit sphere of $\mathbf{R}^d$ which is endowed with the induced topology of $\mathbf{R}^d$.

Suppose that the distribution of X has a continuous density function. If $\|X\| \neq 0$, define the pair (R, Θ) in $\mathbf{R}_+^* \times \mathbf{S}^{d-1}$ by $R = \|X\|$ and $\Theta = \frac{X}{\|X\|}$. For all θ, assume the distribution of R given $\Theta = \theta$ is defined by the continuous conditional distribution function,

$$F_\theta(r) = P\big\{R \leq r \mid \Theta = \theta\big\}. \tag{3.1}$$

Denote by F_θ^{-1} its generalized inverse.

Definition 3.1 For a given u, $0 < u < 1$, the *u-level isobar* from the distribution of (R, Θ) is defined by:

$$\begin{aligned} &\mathbf{S}^{d-1} \to \mathbf{R}_+^*, \\ &\theta \to F_\theta^{-1}(u) = \rho_u(\theta). \end{aligned}$$

The corresponding surface is also called isobar. See Fig. 3.1.

We suppose that for u fixed, the mapping F_θ^{-1} is continuous and strictly positive. So, isobars are closed surfaces included in each other for increasing levels. For bivariate distributions, isobars are classical curves in polar coordinates. Very different shapes of isobars can be considered according to the choice of the distribution.

Let $E_n = (X_1, \ldots, X_n)$ be a sample of independent random variables with the same distribution as X. For each $1 \leq i \leq n$ there is almost surely a unique isobar from the distribution of R given $\Theta = \theta$ which contains (R_i, Θ_i). We define the maximum value in E_n as the point $X_n^* = (R_n^*, \Theta_n^*)$ which corresponds to the upper level isobar. So, $F_{\Theta_n^*}(R_n^*) = \max_{1 \leq i \leq n} U_i$, with $U_i = F_{\Theta_i}(R_i)$.

We call X_n^* the isobar-maximum of $X_1, \ldots, X_n$; see Fig. 3.2.

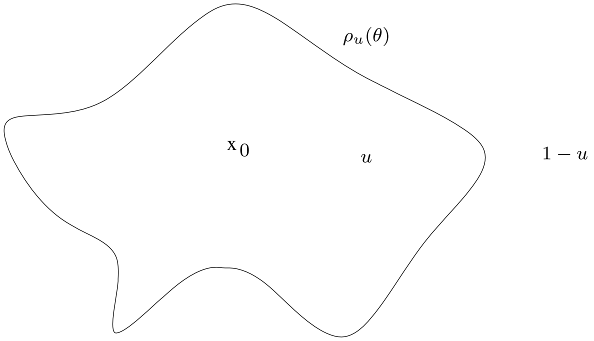


Fig. 3.1 u-level isobar

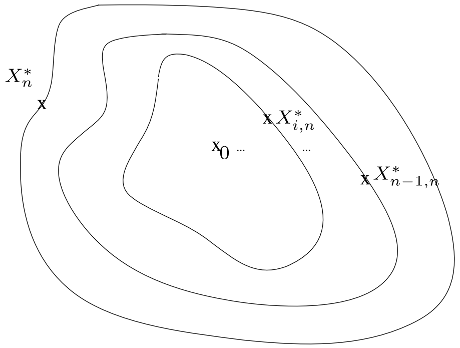


Fig. 3.2 Isobar-maximum

Definition 3.2 The *maximum value* in $X_1, \dots, X_n$ is defined as the point X_n^* which belongs to the upper level surface, i.e., the surface which has a level equal to $\max_{1 \le i \le n} U_i$.

The multivariate sample $X_1, \dots, X_n$ is then ordered according to the increasing levels, $U_{1,n} \le \cdots \le U_{n,n}$, of the corresponding isobar surfaces, following the classical notation for the order statistics of unidimensional samples, and the corresponding *order statistics* are denoted by

$$X_{1,n}^* = \left(R_{1,n}^*, \Theta_{1,n}^*\right), \quad \dots, \quad X_{n,n}^* = \left(R_n^*, \Theta_n^*\right) = X_n^*. \tag{3.2}$$

Obviously, we are not able to find this maximum value of a sample from an unknown distribution, whereas it can be done with the farthest point from the origin or with the fictitious point having the largest coordinates of the sample. However, this kind of extreme value and, more generally, the extreme values obtained by ordering the sample according to the levels, hold more information on the conditional distributions tails and allow a statistic survey of the isobars.[1]

We are well aware that the above definition depends on the underlying distribution and in contexts with just a given data set, it cannot be applied when the data

[1] A paper concerning the estimation of isobars is in progress, Barme-Delcroix and Brito (2011).

generating distribution is not known. This is usually a deficiency but in this contribution, where we want to check if a given distribution is suitable for modeling a data structure, we are able to use this natural notion of ordering since we suppose that the distribution is known.

Remark 1 Note that the maximum value is a sample point and is defined intrinsically, only with the underlying distribution, taking into account the shape of the distribution.

Remark 2 Since for all θ and for all $0 \le r \le 1$, $P(F_\Theta(R) \le r \mid \Theta = \theta) = F_\theta(F_\theta^{-1}(r)) = r$, the variables $U_i = F_{\Theta_i}(R_i)$ are independent and uniformly distributed over [0, 1].

Remark 3 We could imagine a more general way to order the sample. For example, by considering an increasing sequence of Borelians, according to a criterion to define, and not necessarily related to the Euclidean norm. But it is not the purpose of this contribution.

Remark 4 The definition depends of the choice of the origin and the equations of isobars change and then the ordering completely changes if we change the origin. For a given data set one can estimate the origin by using the barycenter of the sample points. But for many practical situations the origin is given in a natural way (for instance, consider a rescue center and the accidents all around).

3.3 Weak Stability of Multivariate Extremes and Outlier-Resistance

In Barme-Delcroix and Gather (2002), we have given a framework and definitions of the terms outlier-proneness and outlier-resistance of multivariate distributions based on our definition of multivariate extreme values. As for the univariate case, Green (1976), Gather and Rauhut (1990), we have classified the multivariate distributions w.r.t. their outlier-resistance and proneness. Characterizations have been provided in terms of the distribution functions. Let us recall the main results. We start with defining the weak stability of the extremes. It has been shown in Delcroix and Jacob (1991) that the conditional distribution of R_n^* given Θ_n^* is F_θ^n, hence the distributions of (R_n^*, Θ_n^*) and (R, Θ) have the same set of isobars which led to the following definition of the weak stability (or stability in probability) of the sequence $(X_n^*)_n$.

Definition 3.3 The sequence $(X_n^*)_n = ((R_n^*, \Theta_n^*))_n$ of the isobar-maxima is called stable in probability if and only if there is a sequence $(g_n)_n$ of isobars satisfying

$$R_n^* - g_n(\Theta_n^*) \xrightarrow{P} 0. \tag{3.3}$$

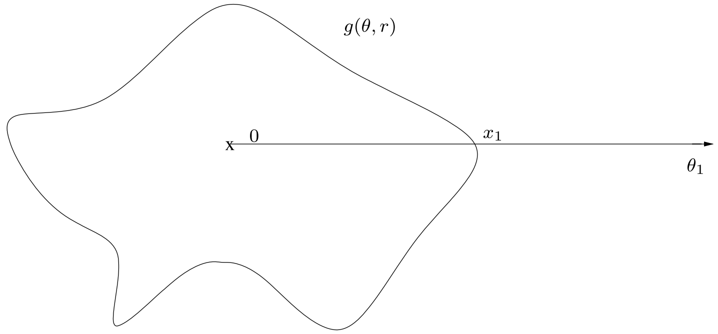


Fig. 3.3 Isobar containing an arbitrarily point $x_1 = (1, \theta_1)$

Following Geffroy (1958), we will see in this section that it is possible to choose $g_n(\theta) = F_\theta^{-1}(1 - \frac{1}{n})$. Examples are given after Theorem 3.2.

We suppose now that F_θ is one-to-one. It is convenient to fix arbitrarily a point $x_1 = (1, \theta_1)$, θ_1 in $\mathbf{S}^{d-1}$. For every point $x = (r, \theta_1)$, there is a unique surface $g(\theta, r)$, θ in $\mathbf{S}^{d-1}$, containing x, which has a level denoted by $u(r)$ and which is given by

$$g(\theta, r) = \rho_{u(r)}(\theta) = {F_\theta}^{-1}\big(F_{\theta_1}(r)\big). \tag{3.4}$$

Note that $g(\theta_1, r) = r$; see Fig. 3.3. Moreover, the mapping $r \to u(r)$ from $\mathbf{R}_+^*$ into $\mathbf{R}_+^*$ is increasing and one-to-one.

The following conditions (H) and (K) will be needed.

(H) There exist $0 < \alpha \leq \beta < \infty$ such that for all θ in $\mathbf{S}^{d-1}$ and for all $r > 0$:

$$\alpha \leq \frac{\partial g}{\partial r}(\theta, r) \leq \beta.$$

(K) For all $\varepsilon > 0$, there exists $\eta > 0$ such that for all $r > 0$:

$$\sup_\theta \big\{g(\theta, r + \eta) - g(\theta, r - \eta)\big\} < \varepsilon.$$

Clearly, (H) implies (K).

Remark 5 Condition (H) entails a regularity property of the isobars following from the mean value theorem:

For all $\beta_0 > 0$, there exists $\eta = \beta_0 \frac{\alpha}{\beta} > 0$ and for all $r > 0$, there exist two isobars $h_{\beta_0}(\theta, r) = g(\theta, r + \frac{\beta_0}{\beta})$ and $\tilde{h}_{\beta_0}(\theta, r) = g(\theta, r - \frac{\beta_0}{\beta})$ such that for all θ,

$$g(\theta, r) - \beta_0 < \tilde{h}_{\beta_0}(\theta, r) < g(\theta, r) - \eta < g(\theta, r) + \eta < h_{\beta_0}(\theta, r) < g(\theta, r) + \beta_0.$$

Note that η does not depend on r.

For all $i \geq 1$, let W_i be the intersection of the half axis $\overrightarrow{0\theta_1}$ containing the point $x_1 = (1, \theta_1)$ and the isobar-surface containing X_i; $W_i = F_{\theta_1}^{-1}(F_{\Theta_i}(R_i))$. See

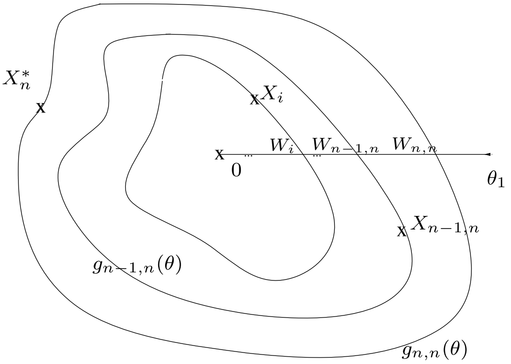


Fig. 3.4 The order statistics of the real sample $W_1, \ldots, W_n$

Fig. 3.4. In fact, $(W_n)_n$ is a sequence of i.i.d. variables from the distribution F_{θ_1}. As usual $W_{1,n} \le \cdots \le W_{n-1,n} \le W_{n,n}$ denotes the corresponding order statistics for the sample $(W_1, \ldots, W_n)$. Let $g_{n,n}$ denote the isobar containing $X_n^* = X_{n,n}$ and $W_{n,n}$, and $g_{n-1,n}$ the isobar containing $X_{n-1,n}$ and $W_{n-1,n}$.

The next theorem ensures that the concept of ordering multivariate data according to the isobar surfaces yields analogous results to the univariate case, Barme-Delcroix and Gather (2002).

Theorem 3.1

1. *Under condition* (K) *the sequence* $(X_n^*)_n$ *is stable in probability if* $(W_{n,n})_n$ *is stable in probability.*
2. *Under condition* (H) *the sequence* $(W_{n,n})_n$ *is stable in probability if and only if* $(X_n^*)_n$ *is stable in probability.*
3. *Consider for some fixed integer* $1 \le \alpha \le n$ *the sequence* $(X_{n-\alpha+1,n})_n$, *this being defined by ordering the sample according to increasing levels by*

$$X_{1,n}, \ldots, X_{n-\alpha+1,n}, \ldots, X_{n,n} = X_n^*.$$

Let (H) *be satisfied. Then* $(X_n^*)_n$ *is stable in probability if and only if* $(X_{n-\alpha+1,n})_n$ *is stable in probability.*

As an application of the weak stability of extreme values of multivariate samples we can now define the notion of Absolute Outlier-Resistance. Recall that Green (1976) called a univariate distribution F absolutely outlier-resistant if for all $\epsilon > 0$:

$$\lim_{n \to +\infty} P(W_{n,n} - W_{n-1,n} > \varepsilon) = 0,$$

where $W_{1,n} \le \cdots \le W_{n-1,n} \le W_{n,n}$ are the usual univariate order statistics of $W_1, \ldots, W_n$, distributed identically according to F.

Following Green (1976), we can now propose the definition of multivariate Absolute Outlier-Resistant distributions.

Definition 3.4 The distribution of the multivariate r.v. (R, Θ) is absolutely outlier-resistant (AOR), if and only if for all θ:

$$g_{n,n}(\theta) - g_{n-1,n}(\theta) \overset{P}{\to} 0. \tag{3.5}$$

For a real sample $W_1, \ldots, W_n$ it has been shown in Geffroy (1958) and Gnedenko (1943), that $(W_{n,n})_n$ is stable in probability if and only if $W_{n,n} - W_{n-1,n} \overset{P}{\to} 0$. The following theorem, Barme-Delcroix and Gather (2002), gives an analogous result and a characterization of weak stability by the tail behaviour of the underlying distribution. Let $\bar{F}_\theta = 1 - F_\theta$.

Theorem 3.2 *Let condition* (H) *be satisfied. All the following statements are equivalent*:

1. *The distribution of* (R, Θ) *is* AOR.
2. $(X_n^*)_n$ *is stable in probability.*
3. *For every fixed integer* $1 \le \alpha \le n$, $(X_{n-\alpha+1,n})_n$ *is stable in probability.*
4. *There exists* θ_1 *such that* $\lim_{x\to+\infty} \bar{F}_{\theta_1}(x)/\bar{F}_{\theta_1}(x-h) = 0,$ *for all* $h > 0$.
5. *For all* θ, $\lim_{x\to+\infty} \bar{F}_{\theta_1}(x)/\bar{F}_{\theta_1}(x-h) = 0,$ *for all* $h > 0$.
6. $W_{n,n} - W_{n-1,n} \overset{P}{\to} 0$.
7. $(W_{n,n})_n$ *is stable in probability.*
8. *For all* θ, *the distribution* F_θ *is* AOR.
9. *There exists* θ_1 *such that the distribution* F_{θ_1} *is* AOR.

Other characterizations can be found in Barme-Delcroix and Gather (2002).

Example 1 In the first example, $F_\theta(r) = (1 - e^{-\alpha(\theta)r^m})1_{\{r>0\}}$, where $m > 0$, and α is a continuous strictly positive function over $[0, 2\pi]$ such that $\alpha(0) = \alpha(2\pi)$. For a fixed θ_1 and for every $r > 0$, the $u(r)$-level isobar $g(\theta, r)$ is defined, according to (3.4), by

$$g(\theta, r) = \left(\frac{\alpha(\theta_1)}{\alpha(\theta)}\right)^{1/m} r,$$

so that (H) is fulfilled. Theorem 3.2(5) shows that $(X_n^*)_n$ is stable in probability if and only if $m > 1$.

Example 2 For a bivariate Gaussian centered distribution with covariance matrix $\left(\begin{smallmatrix}\sigma^2 & 0\\ 0 & \tau^2\end{smallmatrix}\right)$, we have $g(\theta, r) = r\phi(\theta)$ with $\phi(\theta) = \frac{1}{\sqrt{2}\sigma}(\frac{cos^2\theta}{2\sigma^2} + \frac{sin^2\theta}{2\tau^2})^{\frac{-1}{2}}$ and the isobars are the density contours. Note that condition (H) is satisfied. For $\sigma = \tau = 1$ the distribution is spherically symmetric and the isobars are circles. Hence, in this particular case, the ordering of the sample is the ordering of the norms of the sample points. In this example, $F_\theta(r) = 1 - \exp(-r^2\phi(\theta))$. Following Theorem 3.2(5) we conclude that the distribution is AOR.

Similarly, we can define outlier-prone multivariate distributions, that is distributions such that there exist observations far apart from the main group of the data.

Definition 3.5 The distribution of (R, Θ) is called absolutely outlier-prone, (AOP), if and only if for all θ there exist $\varepsilon > 0$, $\delta > 0$ and an integer n_θ, such that for all θ and for all $n \geq n_\theta$:

$$P\big(g_{n,n}(\theta) - g_{n,n-1}(\theta) > \varepsilon\big) > \delta. \tag{3.6}$$

That is, for all θ, the distribution F_θ is AOP.

Theorem 3.3 *Let condition* (H) *be satisfied. All the following statements are equivalent*:

1. *The distribution of* (R, Θ) *is* AOP.
2. *For all* θ, *there exist* $\alpha > 0$, $\beta > 0$ *such that for all* x
$$\frac{\bar{F}_\theta(x+\beta)}{\bar{F}_\theta(x)} \geq \alpha.$$
3. *There exist* θ_0, $\alpha_0 > 0$ *and* $\beta_0 > 0$ *such that for all* x
$$\frac{\bar{F}_{\theta_0}(x+\beta_0)}{\bar{F}_{\theta_0}(x)} \geq \alpha_0.$$
4. *There exists* θ_0 *such that* F_{θ_0} *is* AOP.

See Barme-Delcroix and Gather (2002) for more details.

3.4 Records for a Multidimensional Sequence

Let $\{X_n = (R_n, \Theta_n), n \geq 1\}$ be a sequence of independent, identically distributed random variables as $X = (R, \Theta)$ in the previous sections, with common conditional distribution function $F_\theta(\cdot)$. According to the definitions of Sect. 3.2, we associate the sequence of the levels, that is the sequence of the independent, uniformly distributed over $[0, 1]$ variables $\{U_n = F_{\Theta_n}(R_n), n \geq 1\}$. As usual, Resnick (1973a), Galambos (1987), we can define the notion of record values for the sequence $\{U_n, n \geq 1\}$. U_j is a record value for this sequence if and only if:

$$U_j > \max(U_1, \ldots, U_{j-1}),$$

with the convention that U_1 is a record value.

The indices at which record values occur are given by the random variables $\{L_n, n \geq 0\}$ defined by

$$L_0 = 1,$$

and

$$L_n = \min(j : j > L_{n-1}, U_j > U_{L_{n-1}}).$$

The distribution function for a uniform variable being continuous, the variables L_n are well defined with probability one.

Note that $U_{L_n} = \max(U_1, \dots, U_{L_n})$.

Now we can define the record values for the multidimensional sequence $\{X_n = (R_n, \Theta_n), n \geq 1\}$, since the sequence has been ordered according to the increasing levels.

Definition 3.6 The record values for the sequence $\{X_n = (R_n, \Theta_n), n \geq 1\}$ are defined by:

$$\left\{X_{L_n} = (R_{L_n}, \Theta_{L_n}), n \geq 0\right\}. \tag{3.7}$$

So the definition of the record values for the sequence of the levels $\{U_n, n \geq 1\}$ induces the definition of the record values for the sequence $\{X_n = (R_n, \Theta_n), n \geq 1\}$. The definition seems relevant because it is based on the *probability* to be at a certain distance from the origin, given the angle. Thus, we consider the intrinsic properties of the multivariate distribution.

Lemma 1 *For all $n \geq 0$, The variables Θ_{L_n} and Θ are identically distributed.*

Proof The record value of the sequence $\{X_n, n \geq 1\}$, associated with the record time L_n is almost surely defined as the point X_{L_n} with polar representation

$$(R_{L_n}, \Theta_{L_n}) = \sum_{i=1}^{+\infty} (R_i, \Theta_i) 1_{\mathcal{E}_i}, \tag{3.8}$$

where

$$\mathcal{E}_i = \left\{F_{\Theta_i}(R_i) = U_{L_n} = \max(U_1, \dots, U_{L_n}) = \max_{j=1}^{L_n} F_{\Theta_j}(R_j)\right\}. \tag{3.9}$$

As noticed in Remark 2, $P(F_\Theta(R) \leq r \mid \Theta = \theta) = r$, and for each $j \geq 1$ the variables Θ_j and $U_j = F_{\Theta_j}(R_j)$ are independent. It follows that $\{\Theta_j; j \geq 1\}$ and $\{F_{\Theta_j}(R_j); j \geq 1\}$ are independent. Therefore for each $j \geq 1$, Θ_j and $1_{\mathcal{E}_j}$ are independent, since the variables L_j are $\sigma(U_j)$—measurable. Consequently, for any Borel set C of S^{k-1}:

$$\begin{aligned} P(\Theta_{L_n} \in C) &= P\left(\sum_{i=1}^{+\infty} \Theta_i \, 1_{\mathcal{E}_i} \in C\right) = \sum_{i=1}^{+\infty} P(\Theta_i \in C; \mathcal{E}_i) \\ &= \sum_{i=1}^{+\infty} P(\Theta_i \in C) P(\mathcal{E}_i) = P(\Theta \in C). \end{aligned} \tag{3.10}$$

□

Lemma 2 *Any isobar from the distribution of R given Θ is also an isobar from the distribution of R_{L_n} given Θ_{L_n}.*

Proof Let $g(\theta) = F_\theta^{-1}(u)$ be an u-level isobar from the distribution of R given $\Theta = \theta$ and let $\mathcal{B}$ be the event

$$\mathcal{B} = \left\{R_{L_n} \leq F^{-1}_{\Theta_{L_n}}(u)\right\}.$$

Since $\mathcal{B} = \bigcap_{i=1}^{L_n} \{F_{\Theta_i}(R_i) \leq u\} = \{\max(U_1, \ldots, U_{L_n}) \leq u\}$, $\mathcal{B}$ is independent of $\{\Theta_j, j \geq 1\}$. Thus for any Borel set C of $\mathbf{S}^{d-1}$, (3.10) implies:

$$\begin{aligned} P(\Theta_{L_n} \in C; \mathcal{B}) &= \sum_{i=1}^{+\infty} P(\Theta_i \in C; \mathcal{E}_i; \mathcal{B}) = \sum_{i=1}^{+\infty} P(\Theta_i \in C) P(\mathcal{E}_i; \mathcal{B}) \\ &= P(\Theta_{L_n} \in C) P(\mathcal{B}). \end{aligned}$$

Thus Θ_{L_n} and $\mathbf{1}_{\mathcal{B}}$ are independent; therefore,

$$\begin{aligned} P\big(R_{L_n} \leq F^{-1}_{\Theta_{L_n}}(u) \mid \Theta_{L_n} = \theta\big) = P(\mathcal{B}) &= \sum_{k=1}^{+\infty} P\left(\bigcap_{i=1}^{L_n} F_{\Theta_i}(R_i) \leq u; L_n = k\right) \\ &= \sum_{k=1}^{+\infty} u^k P(L_n = k). \end{aligned}$$

□

3.5 Weak Stability of Multivariate Records

The results of the previous section state that both the distributions of R given Θ and the distributions of R_{L_n} given Θ_{L_n} have the same set of isobars. Hence, we deal only with the formers. In the sequel, any u-level isobar from the distribution of R given Θ is labelled as u-level isobar. So we may give the following definitions.

Definition 3.7 The sequence $(X_{L_n})_n = ((R_{L_n}, \Theta_{L_n}))_n$ of the multidimensional records is stable in probability if and only if there is a sequence $(g_n)_n$ of isobars satisfying

$$R_{L_n} - g_n(\Theta_{L_n}) \xrightarrow{P} 0. \tag{3.11}$$

We can also define the relative stability for the multidimensional records.

Definition 3.8 The sequence $(X_{L_n})_n = ((R_{L_n}, \Theta_{L_n}))_n$ of the multidimensional records is relatively stable in probability if and only if there is a sequence $(g_n)_n$ of isobars satisfying

$$\frac{R_{L_n}}{g_n(\Theta_{L_n})} \xrightarrow{P} 1. \tag{3.12}$$

As in Sect. 3.3, *we suppose that F_θ is one-to-one.* In the next theorem, it is shown that the weak stability of the sequence of the multidimensional records $(X_{L_n})_n = ((R_{L_n}, \Theta_{L_n}))_n$ can be investigated through the stability of the real sequence $(W_{L_n})_n$. See Fig. 3.5. The conditions (H) and (K) will be useful again.

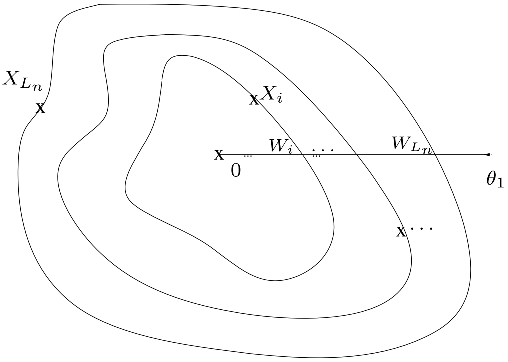


Fig. 3.5 The sequence of records

Theorem 3.4

1. *Under condition* (K) *the sequence* $(X_{L_n})_n$ *is stable in probability if the sequence* $(W_{L_n})_n$ *is stable in probability.*
2. *Under condition* (H) *the sequence* $(X_{L_n})_n$ *is stable in probability if and only if the sequence* $(W_{L_n})_n$ *is stable in probability.*

Proof (1) If $(W_{L_n})_n$ is stable in probability, then there exists a sequence (w_n) such that $W_{L_n} - w_n \overset{P}{\to} 0$. According to (K), for $\epsilon > 0$ there exists $\eta > 0$ such that $\sup_\theta\{g(\theta, r+\eta) - g(\theta, r-\eta)\} < \varepsilon$, for all $w > 0$. Let $h_n^\eta(\theta) = g(\theta, w_n + \eta)$ and $h_n^{-\eta}(\theta) = g(\theta, w_n - \eta)$ and put $g(\theta, w_n) = h_n(\theta)$. We have therefore

$$\begin{aligned}\left\{|W_{L_n} - w_n| \le \eta\right\} &= \left\{h_n^{-\eta}(\theta_1) \le W_{L_n} \le h_n^\eta(\theta_1)\right\}\\ &\subset \left\{h_n^{-\eta}(\Theta_{L_n}) \le R_{L_n} \le h_n^\eta(\Theta_{L_n})\right\}\\ &\subset \left\{\left|R_{L_n} - h_n(\Theta_{L_n})\right| \le \epsilon\right\}\end{aligned}$$

implying that $R_{L_n} - h_n(\Theta_{L_n}) \overset{P}{\to} 0$.

(2) Conversely, if there exists a sequence of surfaces g_n such that $R_{L_n} - g_n(\Theta_{L_n}) \overset{P}{\to} 0$, denote by w_n the intersection of the half axis $0\theta_1$ with g_n. According to (H), there exist α and β such that

$$g(\theta, w_n) + \lambda\alpha \le g(\theta, w_n + \lambda) \le g(\theta, w_n) + \lambda\beta$$

and

$$g(\theta, w_n) - \lambda\beta \le g(\theta, w_n - \lambda) \le g(\theta, w_n) - \lambda\alpha$$

for all $\lambda > 0$ and all θ. Given $\epsilon > 0$, it is possible to choose $\lambda = \epsilon/\beta$ and $\eta = \epsilon\alpha/\beta$ and to take

$$\begin{aligned}h_n(\theta) &= g(\theta, w_n + \lambda),\\ \tilde{h}_n(\theta) &= g(\theta, w_n - \lambda).\end{aligned}$$

It follows that

$$\left\{\left|R_{L_n} - g_n(\Theta_{L_n})\right| \le \eta\right\} \subset \left\{\tilde{h}_n(\Theta_{L_n}) \le R_{L_n} \le h_n(\Theta_{L_n})\right\} \subset \left\{|W_{L_n} - w_n| \le \epsilon\right\},$$

which completes the proof. □

Now we can use unidimensional criteria to obtain characterizations for the weak stability or relative stability of multidimensional records. Following Resnick (1973a,b), let us define for all θ and for all $r > 0$, the integrated hazard function

$$\mathcal{R}_\theta(r) = -\log\bigl(1 - F_\theta(r)\bigr).$$

Theorem 3.5 *Under condition* (H), *the sequence* $(X_{L_n})_n$ *is stable in probability if and only if*

$$R_{L_n} - \mathcal{R}^{-1}_{\Theta_{L_n}}(n) \xrightarrow{P} 0. \tag{3.13}$$

Or, equivalently, if and only if there exists θ_1 *such that for all* $\epsilon > 0$,

$$\lim_{r\to+\infty} \frac{\mathcal{R}_{\theta_1}(r+\epsilon) - \mathcal{R}_{\theta_1}(r)}{\mathcal{R}^{1/2}_{\theta_1}(r+\epsilon)} = +\infty. \tag{3.14}$$

Or, equivalently, if and only if for all θ *and for all* $\epsilon > 0$,

$$\lim_{r\to+\infty} \frac{\mathcal{R}_{\theta}(r+\epsilon) - \mathcal{R}_{\theta}(r)}{\mathcal{R}^{1/2}_{\theta}(r+\epsilon)} = +\infty. \tag{3.15}$$

Theorem 3.6 *Under condition* (H), *the sequence* $(X_{L_n})_n$ *is relatively stable in probability if and only if*

$$\frac{R_{L_n}}{\mathcal{R}^{-1}_{\Theta_{L_n}}(n)} \xrightarrow{P} 1. \tag{3.16}$$

Or, equivalently, if and only if there exists θ_1 *such that for all* $k > 1$,

$$\lim_{r\to+\infty} \frac{\mathcal{R}_{\theta_1}(kr) - \mathcal{R}_{\theta_1}(r)}{\mathcal{R}^{1/2}_{\theta_1}(kr)} = +\infty. \tag{3.17}$$

Or, equivalently, if and only if for all θ *and for all* $k > 1$,

$$\lim_{r\to+\infty} \frac{\mathcal{R}_{\theta}(kr) - \mathcal{R}_{\theta}(r)}{\mathcal{R}^{1/2}_{\theta}(kr)} = +\infty. \tag{3.18}$$

Remark 6 These theorems imply that a convenient sequence of isobars satisfying the conditions (3.11) and (3.12) of Definitions 3.7 and 3.8 is given by $g_n(\theta) = \mathcal{R}_\theta^{-1}(n) = F_\theta^{-1}(1 - \exp(-n))$.

Example 3 Recall that in the first example, $F_\theta(r) = (1 - e^{-\alpha(\theta)r^m})1_{\{r>0\}}$, where $m > 0$, and α is a continuous strictly positive function over $[0, 2\pi]$ such that $\alpha(0) =$

$\alpha(2\pi)$. For a fixed θ_1 and for every $r > 0$, the $u(r)$-level isobar $g(\theta, r)$ is defined, according to (3.4), by

$$g(\theta, r) = \left(\frac{\alpha(\theta_1)}{\alpha(\theta)}\right)^{1/m} r,$$

and (H) is fulfilled. In this case $\mathcal{R}_\theta(r) = \alpha(\theta) r^m$; so condition (3.14) or (3.15) of Theorem 3.5 is satisfied for $m > 2$ and the sequence $(X_{L_n})_n$ is stable in probability for $m > 2$. Moreover, for all $m > 0$, condition (3.17) or (3.18) is satisfied and the sequence $(X_{L_n})_n$ is relatively stable in probability for all $m > 0$.

Example 4 For a bivariate Gaussian centered distribution with covariance matrix $\left(\begin{smallmatrix}\sigma^2 & 0\\ 0 & \tau^2\end{smallmatrix}\right)$, we have $g(\theta, r) = r\phi(\theta)$ with $\phi(\theta) = \frac{1}{\sqrt{2}\sigma}\left(\frac{\cos^2\theta}{2\sigma^2} + \frac{\sin^2\theta}{2\tau^2}\right)^{\frac{-1}{2}}$. We know already that condition (H) is satisfied. In this example, $F_\theta(r) = 1 - \exp(-r^2\phi(\theta))$ and $\mathcal{R}_\theta(r) = r^2\phi(\theta)$ and we can easily check the conditions of Theorem 3.5 and conclude that the sequence $(X_{L_n})_n$ is stable in probability.

3.6 Conclusions

We have shown that, by using the isobar surfaces approach, the multivariate weak stability properties for the extreme values and record values may be investigated in a univariate way. We could now focus, in a future work, on finding characterizations of the multivariate a.s. stability of the record values as it has been done for the intermediate order statistics in Barme-Delcroix and Brito (2001).

References

Abdous, B., & Theodorescu, R. (1992). Note on the spatial quantile of a random vector. *Statistics & Probability Letters*, *13*, 333–336.

Barme-Delcroix, M. F. (1993). Localisation asymptotique des échantillonsmultidimensionnels. *Bulletin de la Société Mathématique de Belgique. Série B*, *45*, 2.

Barme-Delcroix, M. F., & Brito, M. (2001). Multivariate stability and strong limiting behaviour of intermediate order statistics. *Journal of Multivariate Analysis*, *79*, 157–170.

Barme-Delcroix, M. F., & Brito, M. (2011). Quantile estimation for a bivariate sample. In *Proceedings of the 7th conference on extreme value analysis*, Lyon.

Barme-Delcroix, M. F., & Gather, U. (2002). An isobar surfaces approach to multidimensional outlier-proneness. *Extremes*, *5*, 131–144.

Barme-Delcroix, M. F., & Gather, U. (2007). Limit laws for multidimensional extremes. *Statistics & Probability Letters*, *77*, 1750–1755.

Barnett, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistical Society. Series A. General*, *139*, 318–355.

Berlinet, A., Gannoun, A., & Matzner-Løber, E. (2001). Asymptotic normality of convergent estimates of conditional quantiles. *Statistics*, *35*, 139–169.

De Haan, L., & Huang, X. (1995). Large quantile estimation in a multivariate setting. *Journal of Multivariate Analysis*, *53*, 247–263.

Delcroix, M. F., & Jacob, P. (1991). Stability of extreme value for a multidimensional sample. *Statistique Et Analyse Des Données*, *16*, 1–21.

Einmahl, H. J., & Mason, D. M. (1992). Generalized quantile processes. *The Annals of Statistics*, *20*, 1062–1078.

Falk, M., & Reiss, R. D. (2005). On Pickands coordinates in arbitrary dimensions. *Journal of Multivariate Analysis*, *92*, 426–453.

Galambos, J. (1987). *The asymptotic theory of extreme order statistics*. Malabar: Krieger.

Gather, U., & Rauhut, B. (1990). Outlier behaviour of probability distributions. *Journal of Statistical Planning and Inference*, *26*, 237–252.

Geffroy, J. (1958). Contribution à la théorie des valeurs extrêmes. *Publications de L'Institut de Statistique de L'Université de Paris*, *7*, 37–185.

Geffroy, J. (1961). Localisation asymptotique du polyèdre d'appui d'un échantillon Laplacien à k dimension. *Publications de L'Institut de Statistique de L'Université de Paris*, *10*, 213–228.

Gnedenko, B. V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, *44*, 423–453.

Green, R. F. (1976). Outlier-prone and outlier-resistant distributions. *Journal of the American Statistical Association*, *71*, 502–505.

Hallin, M., Paindaveine, D., & Šiman, M. (2010). Multivariate quantiles and multiple-output regression quantiles: from L_1 optimization to halfspace depth. *The Annals of Statistics*, *38*, 635–669.

Heffernan, J. E., & Tawn, J. A. (2004). A conditional approach for multivariate extreme values. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *66*, 497–546.

Ivanková, K. (2010). *Isobars and the efficient market hypothesis*. Working paper, Institute of Economic Studies, Faculty of Social Sciences, Charles University in Prague.

Maller, R. R. (1990). Defining extremes and trimming by minimum covering sets. *Stochastic Processes and Their Applications*, *35*, 169–180.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, *1*, 327–332.

Resnick, S. I. (1973a). Limit laws for record values. *Stochastic Processes and Their Applications*, *1*, 67–82.

Resnick, S. I. (1973b). Records values and maxima. *Annals of Probability*, *1*, 650–662.

Serfling, R. (2002). Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica*, *56*, 214–232.

Chapter 4
High-Breakdown Estimators of Multivariate Location and Scatter

Peter Rousseeuw and Mia Hubert

4.1 Introduction

In the multivariate setting, we assume that we have p-dimensional column observations $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$, drawn from a p-variate random variable $X = (X_1, \ldots, X_p)'$ with distribution F on $\mathbb{R}^p$.

In this chapter, we will assume that the distribution F of the uncontaminated data is *elliptical*. We say that F is elliptical if there exists a vector $\boldsymbol{\mu}$, a positive definite matrix Σ, and a nonnegative strictly decreasing real function h such that the density of F can be written in the form

$$f(\mathbf{x}) = \frac{1}{\sqrt{|\Sigma|}} h\big(d^2(\mathbf{x}, \boldsymbol{\mu}, \Sigma)\big) \tag{4.1}$$

in which the *statistical distance* $d(\mathbf{x}, \boldsymbol{\mu}, \Sigma)$ is given by

$$d(\mathbf{x}, \boldsymbol{\mu}, \Sigma) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}. \tag{4.2}$$

The matrix Σ is often called the *scatter* matrix. The multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$ is a specific example of (4.1) with $h(t) = (2\pi)^{-p/2} e^{-t/2}$. Another example is the elliptical p-variate Student distribution with ν degrees of freedom ($0 < \nu < \infty$) for which $h(t) = c_t/(t + \nu)^{(p+\nu)/2}$ with c_t a constant. The case $\nu = 1$ is called the multivariate Cauchy distribution.

Elliptical distributions have the following properties:

1. The contours of constant density are ellipsoids.
2. If the mean and variances of X exist, then $E_F[X] = \boldsymbol{\mu}$ and $\mathrm{Cov}_F[X] = c\Sigma$ with c a constant. (For normal distributions $c = 1$.)

P. Rousseeuw (✉) · M. Hubert
Department of Mathematics, KU Leuven, Celestijnenlaan 200B, 3001 Heverlee, Belgium
e-mail: peter.rousseeuw@wis.kuleuven.be

M. Hubert
e-mail: mia.hubert@wis.kuleuven.be

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_4, © Springer-Verlag Berlin Heidelberg 2013

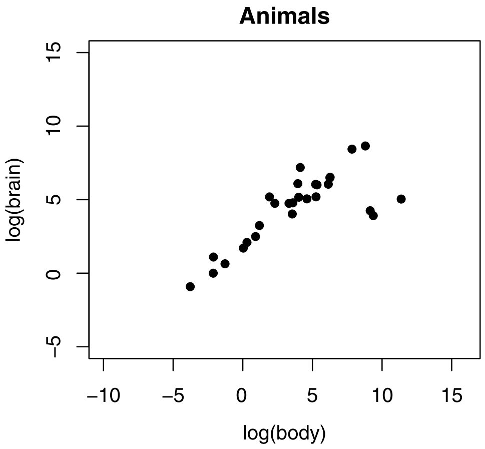


Fig. 4.1 Scatterplot of the logarithm of body and brain weight of 28 animals

3. For any nonsingular matrix A and vector $\mathbf{b}$ the distribution of $AX + \mathbf{b}$ is also elliptical, with mean $A\boldsymbol{\mu} + \mathbf{b}$ and scatter matrix $A\Sigma A'$.
4. The random variable X can be written as

$$X = AZ + \boldsymbol{\mu} \tag{4.3}$$

with A such that $\Sigma = AA'$, and Z a random variable with a spherical distribution, i.e., $f(\mathbf{z}) = h(\|\mathbf{z}\|^2)$.

Example Let us consider the animals data set (Rousseeuw and Leroy 1987), available in R within the MASS package, which contains the body and brain weight of 28 animals. A scatterplot of the logarithm of the observed values (Fig. 4.1) shows that the majority of the data follows an elliptical distribution with a positive correlation, whereas three animals, with large body weight, have a much smaller brain weight than expected under this model. They correspond to the dinosaurs in the data set. We will return to these data in later sections.

Section 4.2 reviews the classical estimators of location and scatter as well as the notion of breakdown value. In Sect. 4.3, we describe multivariate M-estimators and discuss their robustness properties. Section 4.4 is devoted to the highly robust MCD estimator, followed by Sect. 4.5 which describes several other high-breakdown and affine equivariant robust estimators. Some robust but non affine equivariant estimators are summarized in Sect. 4.6, and Sect. 4.7 concludes.

4.2 Classical Estimators

One of the goals of multivariate analysis is to estimate the parameters $\boldsymbol{\mu}$ and Σ from a sample $X_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$. (The sample is thus a matrix with n rows and p

columns.) Under the assumption of multivariate normality, the MLE estimates of $\boldsymbol{\mu}$ and Σ are the sample mean and the (biased) sample covariance matrix:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \quad \text{and} \quad C(X_n) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'.$$

Note that $C(X_n)$ is biased, but $S(X_n) = \frac{n}{n-1}C(X_n)$ is unbiased.

The estimators $\bar{\mathbf{X}}$ and $C(X_n)$ have an important property known as *affine equivariance*. In general, a pair of multivariate location and covariance estimators $(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ is called affine equivariant if for any nonsingular matrix A and any constant vector $\mathbf{b}$ it holds that

$$\hat{\boldsymbol{\mu}}(AX + \mathbf{b}) = A\hat{\boldsymbol{\mu}}(X) + \mathbf{b} \quad \text{and} \quad \hat{\Sigma}(AX + \mathbf{b}) = A\hat{\Sigma}(X)A'. \tag{4.4}$$

Affine equivariance implies that the estimator transforms well under any nonsingular reparametrization of the space of the $\mathbf{x}_i$. The data might for instance be rotated, translated, or rescaled (for example through a change of the measurement units). Note that a transformation AX of the variables $(X_1, \dots, X_p)'$ corresponds to the matrix product $X_n A'$.

At the normal model the sample mean and the sample covariance matrix are consistent and asymptotically normal with maximal efficiency, but they lack robustness:

- Their influence function (Hampel et al. 1986) is unbounded. Let $T(F)$ be the statistical functional mapping a distribution F to its mean $\boldsymbol{\mu}_F = \mathrm{E}_F[X]$ and let $V(F)$ be the statistical functional mapping F to its covariance matrix $\Sigma_F = \mathrm{Cov}_F[X]$. Then at any $\mathbf{z} \in \mathbb{R}^p$

$$\begin{aligned}\mathrm{IF}(\mathbf{z}; T, F) &= \mathbf{z} - \boldsymbol{\mu}_F,\\ \mathrm{IF}(\mathbf{z}; V, F) &= (\mathbf{z} - \boldsymbol{\mu}_F)(\mathbf{z} - \boldsymbol{\mu}_F)' - \Sigma_F.\end{aligned}$$

  Clearly, both influence functions are unbounded in $\mathbf{z}$.
- The asymptotic breakdown value of the mean is zero. More precisely, let $F_{\varepsilon,H} = (1-\varepsilon)F + \varepsilon H$ for any distribution H, then the asymptotic breakdown value is defined as

$$\varepsilon^*(T, F) = \inf\Big\{\varepsilon > 0 : \sup_H \big\|T(F_{\varepsilon,H})\big\| = \infty\Big\}. \tag{4.5}$$

- The asymptotic breakdown value of the classical covariance matrix is zero too. Let us denote the eigenvalues of a p.s.d. matrix by $\lambda_1 \geqslant \cdots \geqslant \lambda_p \geqslant 0$. Then we define the implosion breakdown value of a scatter functional V at F as

$$\varepsilon^*_{\text{imp}}(V, F) = \inf\Big\{\varepsilon > 0 : \inf_H\big\{\lambda_p\big(V(F_{\varepsilon,H})\big)\big\} = 0\Big\}$$

  and the explosion breakdown value as

$$\varepsilon^*_{\text{exp}}(V, F) = \inf\Big\{\varepsilon > 0 : \sup_H\big\{\lambda_1\big(V(F_{\varepsilon,H})\big)\big\} = +\infty\Big\}.$$

The breakdown value then equals the smallest of these:

$$\varepsilon^*(V,F) = \min\bigl(\varepsilon^*_{\text{exp}}(V,F), \varepsilon^*_{\text{imp}}(V,F)\bigr).$$

For the classical covariance, the explosion breakdown value is zero.

The *finite-sample* breakdown value of estimators of location and scatter can be defined accordingly. It can be proved (Lopuhaä and Rousseeuw 1991) that any affine equivariant location estimator $\hat{\boldsymbol{\mu}}$ satisfies

$$\varepsilon^*_n(\hat{\boldsymbol{\mu}}, X_n) \leqslant \frac{1}{n}\left\lfloor \frac{n+1}{2} \right\rfloor. \tag{4.6}$$

Any affine equivariant scatter estimator $\hat{\Sigma}$ satisfies the sharp bound (Davies 1987)

$$\varepsilon^*_n(\hat{\Sigma}, X_n) \leqslant \frac{1}{n}\left\lfloor \frac{n-p+1}{2} \right\rfloor \tag{4.7}$$

if the original sample (before contamination) is in *general position*, which means that no hyperplane contains more than p points. At samples that are not in general position, the upper bound in (4.7) is lower and depends on the maximal number of observations on a hyperplane. For affine equivariant location and scatter estimators the asymptotic breakdown value is always at most 0.5. For recent discussions about the breakdown value and equivariance, see Davies and Gather (2005) and the contribution by Müller, Chap. 5.

Example For the animals data set mentioned before, the classical estimates are $\bar{\mathbf{x}} = (3.77,\ 4.43)'$ and

$$S = \begin{pmatrix} 14.22 & 7.05 \\ 7.05 & 5.76 \end{pmatrix},$$

which yields an estimated correlation of $r = 7.05/\sqrt{14.22 \times 5.76} = 0.78$. To visualize these estimates, we can plot the resulting 97.5 % tolerance ellipse in Fig. 4.2. Its boundary consists of points with constant Mahalanobis distance to the mean. In general dimension p, the tolerance ellipsoid is defined as

$$\left\{\mathbf{x};\, \mathrm{MD}(\mathbf{x}) \leqslant \sqrt{\chi^2_{p,0.975}}\right\} \tag{4.8}$$

with $\mathrm{MD}(\mathbf{x}) = d_i(\mathbf{x}, \bar{\mathbf{x}}, S)$ following (4.2). We expect (for large n) that about 97.5 % of the observations belong to this ellipsoid. One could flag an observation as an outlier if it does not belong to the classical tolerance ellipsoid, but in Fig. 4.2 we see that the three dinosaurs do not stand out relative to the ellipse. This is because the outliers have attracted $\bar{\mathbf{x}}$ and, more importantly, have affected the matrix S in such a way that the ellipse has been inflated in their direction.

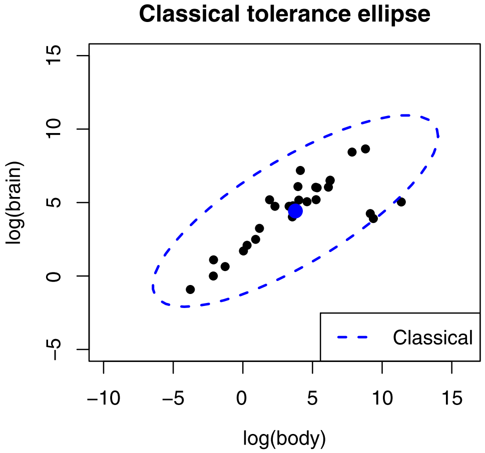


Fig. 4.2 Classical tolerance ellipse of the animals data set

4.3 Multivariate M-Estimators

M-estimators of multivariate location and scatter $(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ were defined in Maronna (1976) as the solution of

$$\sum_{i=1}^{n} W_1\big(d_i^2\big)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}) = \mathbf{0}, \tag{4.9}$$

$$\frac{1}{n}\sum_{i=1}^{n} W_2\big(d_i^2\big)(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})' = \hat{\Sigma} \tag{4.10}$$

with $d_i = d(\mathbf{x}_i, \hat{\boldsymbol{\mu}}, \hat{\Sigma})$ as in (4.2). Note that $\hat{\boldsymbol{\mu}}$ should be a real vector and $\hat{\Sigma}$ a symmetric positive definite matrix. The functions $W_1(t)$ and $W_2(t)$ are real-valued and defined for all $t \geqslant 0$.

If we define $\theta = (\boldsymbol{\mu}, \Sigma)$, $\Psi = (\Psi_1, \Psi_2)$ and

$$\Psi_1(\mathbf{x}, \theta) = W_1\big(d^2\big)(\mathbf{x} - \boldsymbol{\mu}),$$

$$\Psi_2(\mathbf{x}, \theta) = W_2\big(d^2\big)(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' - \Sigma,$$

(4.9) and (4.10) combine into the single M-equation

$$\frac{1}{n}\sum_{i=1}^{n} \Psi(\mathbf{x}_i, \hat{\theta}) = \mathbf{0}.$$

Examples:

- If $W_1(d^2) = W_2(d^2) = 1$, we find the sample mean and sample covariance matrix.

- For an elliptical density (4.1), one finds that the MLE estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ minimize
$$n \log|\hat{\Sigma}| - 2\sum_{i=1}^{n} \log h\big(d_i^2\big).$$
Differentiating with respect to $\boldsymbol{\mu}$ and Σ^{-1} yields the system of equations in (4.9) and (4.10) with $W_1(d^2) = W_2(d^2) = -2h'(d^2)/h(d^2)$. Any MLE is thus an M-estimator.
- In Maronna (1976) several conditions on W_1 and W_2 are given to ensure the existence, uniqueness and consistency of the estimators. Sufficient conditions are that $\sqrt{t}W_1(t)$ and $tW_2(t)$ are bounded, and that $tW_2(t)$ is nondecreasing. A multivariate M-estimator which satisfies the latter condition is called *monotone*. Otherwise it is called redescending.

Some properties:

- Multivariate M-estimators are affine equivariant.
- For a monotone M-estimator (4.9) and (4.10) have a unique solution. This solution can be found by a reweighting algorithm:
  1. Start with initial values $\boldsymbol{\mu}_0$ and Σ_0 such as the coordinatewise median and the diagonal matrix with the squared coordinatewise MAD at the diagonal.
  2. At iteration k compute $d_{ki} = d(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k)$ and
$$\hat{\boldsymbol{\mu}}_{k+1} = \frac{\sum_{i=1}^{n} W_1(d_{ki}^2)\mathbf{x}_i}{\sum_{i=1}^{n} W_1(d_{ki}^2)},$$
$$\hat{\Sigma}_{k+1} = \frac{1}{n}\sum_{i=1}^{n} W_2\big(d_{ki}^2\big)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{k+1})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{k+1})'.$$

  For a monotone M-estimator this algorithm always converges to the unique solution, no matter the choice of the initial values. For a redescending M-estimator the algorithm can convergence to a bad solution, so the initial values are more important in that case.
- Under some regularity conditions on W_1 and W_2, multivariate M-estimators are asymptotically normal.
- The influence function is bounded if $tW_2(t)$ and $\sqrt{t}W_1(t)$ are bounded.
- The asymptotic breakdown value of a monotone M-estimator satisfies
$$\varepsilon^* \leqslant \frac{1}{p+1}.$$
Although monotone M-estimators attain the optimal value of 0.5 in the univariate case, this is no longer true in higher dimensions.

This reveals the main weakness of M-estimators in high dimensions: monotone M-estimators, while computationally attractive, have a rather low breakdown value.

Redescending M-estimators can have a larger breakdown value, but are impractical to compute.

4.4 Minimum Covariance Determinant Estimator

4.4.1 Definition and Properties

The Minimum Covariance Determinant (MCD) estimator (Rousseeuw 1984) was one of the first affine equivariant and highly robust estimators of multivariate location and scatter. Given the parameter h with $[(n+p+1)/2] \leqslant h \leqslant n$, the *raw MCD* is defined as $(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ where:

1. $\hat{\boldsymbol{\mu}}$ is the mean of the h observations for which the determinant of the sample covariance matrix is minimal.
2. $\hat{\Sigma}$ is the corresponding covariance matrix multiplied by a consistency factor.

Note that the MCD estimator can only be computed when $h > p$, otherwise the covariance matrix of any h-subset will be singular. Since $h \geqslant [(n+2)/2]$, this condition is certainly satisfied when $n \geqslant 2p$. To avoid excessive noise, it is however recommended that $n > 5p$. To obtain consistency at the normal distribution, we can use the consistency factor $\alpha / F_{\chi^2_{p+2}}(\chi^2_p, \alpha)$ with $\alpha = h/n$ (see Croux and Haesbroeck (1999)). For small n, we can multiply by an additional finite-sample correction factor given in Pison et al. (2002).

The parameter h controls the breakdown value. At samples in general position $\varepsilon^* = \min(n-h+1, h-p)/n$. The maximal breakdown value (4.7) is achieved by setting $h = [(n+p+1)/2]$. The MCD estimator with $h = [(n+p+1)/2]$ is thus very robust, but unfortunately suffers from a low efficiency at the normal model. For example, the asymptotic relative efficiency of the diagonal elements of the MCD scatter matrix with respect to the sample covariance matrix is only 6 % when $p = 2$, and 20.5 % when $p = 10$. This efficiency can be increased by considering a larger h such as $h \approx 0.75n$. This yields relative efficiencies of 26.2 % for $p = 2$ and 45.9 % for $p = 10$. On the other hand, this choice of h decreases the robustness towards possible outliers.

In order to increase the efficiency while retaining high robustness, one can add a weighting step (Rousseeuw and Leroy 1987; Lopuhaä and Rousseeuw 1991), leading to the *reweighted MCD estimator* (RMCD):

3. Compute $d_i = d(\mathbf{x}_i, \hat{\boldsymbol{\mu}}, \hat{\Sigma})$. Next let

$$\hat{\boldsymbol{\mu}}_{\text{RMCD}} = \frac{\sum_{i=1}^n W(d_i^2)\mathbf{x}_i}{\sum_{i=1}^n W(d_i^2)},$$

$$\hat{\Sigma}_{\text{RMCD}} = \left(\sum_{i=1}^n W\big(d_i^2\big)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{RMCD}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{RMCD}})' \right) \Big/ \left(\sum_{i=1}^n W\big(d_i^2\big) - 1 \right)$$

for some weight function $W(t)$. This weight function should be nonincreasing, bounded, and equal to zero from a certain value t on.

Note the similarity with (4.9) and (4.10). The RMCD estimator can be seen as a one-step M-estimator with the raw MCD estimates as initial starting values. RMCD combines the breakdown value of the original MCD with the better efficiency of M-estimators. A simple yet effective choice for W is to assume that the h selected observations are approximately normally distributed hence the distribution of their d_i^2 is close to χ_p^2, which leads to $W(d^2) = I(d^2 \leqslant \chi^2_{p,\beta})$. This is also the default choice in the `CovMcd` function in the R package `rrcov` (with $\beta = 0.975$). If we take $h \approx 0.5n$ this reweighting step increases the efficiency to 45.5 % for $p = 2$ and to 82 % for $p = 10$.

The MCD estimator is affine equivariant. This property follows from the fact that for each subset of size h, denoted as X_H, the determinant of the covariance matrix of the transformed data equals

$$\bigl|C\bigl(X_H A'\bigr)\bigr| = \bigl|AC(X_H)A'\bigr| = |A|^2\bigl|C(X_H)\bigr|. \tag{4.11}$$

Hence, the optimal h-subset (which minimizes $|C(X_H A')|$) remains the same as for the original data (which minimizes $|C(X_H)|$), and its covariance matrix is transformed appropriately. Afterward, the affine equivariance of the raw MCD location estimator follows from the equivariance of the sample mean. Finally, we note that the robust distances $d_i = d(\mathbf{x}_i, \hat{\boldsymbol{\mu}}, \hat{\Sigma})$ are affine invariant, which implies that the reweighted estimator is equivariant.

The influence functions of the MCD location vector and scatter matrix are bounded (Croux and Haesbroeck 1999; Cator and Lopuhaä 2012). At the standard Gaussian distribution, the influence function of the MCD location estimator becomes zero for all $\mathbf{x}$ with $\|\mathbf{x}\|^2 > \chi^2_{p,\alpha}$, hence large outliers do not influence the estimates. The same happens with the off-diagonal elements of the MCD scatter estimator. On the other hand, the influence function of the diagonal elements remains constant (different from zero) when $\|\mathbf{x}\|^2$ is sufficiently large. Therefore, the outliers still have a bounded influence of the estimator. All these influence functions are smooth, except at those $\mathbf{x}$ with $\|\mathbf{x}\|^2 = \chi^2_{p,\alpha}$. The reweighted MCD estimator has an additional jump in $\|\mathbf{x}\|^2 = \chi^2_{p,0.975}$ due to the discontinuity of the weight function.

Highly robust estimators such as MCD are often used to identify outliers, see, e.g., Rousseeuw and van Zomeren (1990) and Becker and Gather (1999, 2001).

Example For the animals data, the RMCD estimates (for $\alpha = 0.5$) are $\hat{\boldsymbol{\mu}}_{\text{RMCD}} = (3.03, 4.28)'$ and

$$\hat{\Sigma}_{\text{RMCD}} = \begin{pmatrix} 18.86 & 14.16 \\ 14.16 & 11.03 \end{pmatrix}$$

yielding a robust correlation estimate of 0.98 which is higher and corresponds to the correlation of the ‘good’ (non-outlying) data. The corresponding robust tolerance ellipse, now defined as in (4.8) but with the Mahalanobis distances replaced by the robust distances $d_i = d(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_{\text{RMCD}}, \hat{\Sigma}_{\text{RMCD}})$, correctly flags the outliers in Fig. 4.3.

Fig. 4.3 Classical and robust tolerance ellipse of the animals data set

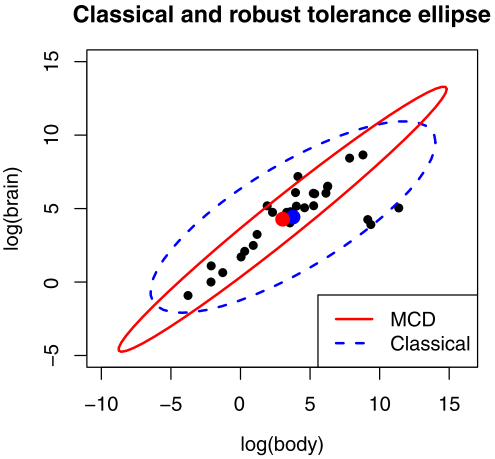
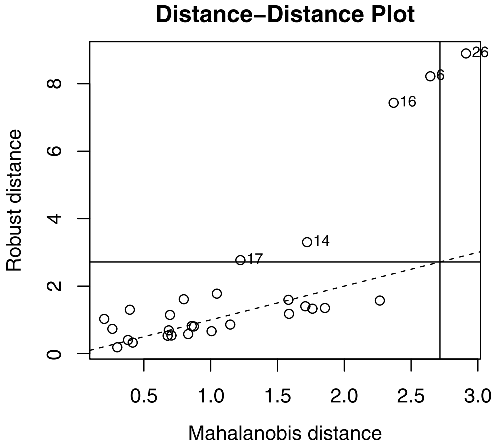


Fig. 4.4 Distance–distance plot of the animals data set



In dimensions $p > 3$, we cannot draw a scatterplot or a tolerance ellipsoid. To explore the differences between the classical and the robust analysis, we can then still draw a *distance-distance plot* which plots the robust distances versus the Mahalanobis distances as in Fig. 4.4. This plot reveals the differences between both methods at a glance, as the three dinosaurs lie far from the dotted line. Note that observation 14, which is little bit outlying, is the human species.

4.4.2 Computation

The exact MCD estimator is very hard to compute, as it requires the evaluation of all $\binom{n}{h}$ subsets of size h. Therefore, one switches to an approximate algorithm such as the FAST-MCD algorithm of Rousseeuw and Van Driessen (1999) which is quite efficient. The key component of the algorithm is the so-called C-step:

Theorem 4.1 *Take $X_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ and let $H_1 \subset \{1, \ldots, n\}$ be subset of size h. Put $\hat{\boldsymbol{\mu}}_1$ and $\hat{\Sigma}_1$ the empirical mean and covariance matrix of the data in H_1. If $|\hat{\Sigma}_1| \neq 0$ define the relative distances $d_i^1 := d(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_1, \hat{\Sigma}_1)$ for all $i = 1, \ldots, n$.*

Now take H_2 such that $\{d_i^1; i \in H_2\} := \{(d^1)_{1:n}, \ldots, (d^1)_{h:n}\}$ where $(d^1)_{1:n} \leqslant (d^1)_{2:n} \leqslant \cdots \leqslant (d^1)_{n:n}$ are the ordered distances, and compute $\hat{\boldsymbol{\mu}}_2$ and $\hat{\Sigma}_2$ based on H_2. Then

$$|\hat{\Sigma}_2| \leqslant |\hat{\Sigma}_1|$$

with equality if and only if $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_1$ and $\hat{\Sigma}_2 = \hat{\Sigma}_1$.

If $|\hat{\Sigma}_1| > 0$, the C-step thus easily yields a new h-subset with lower covariance determinant. Note that the C stands for ‘concentration’ since $\hat{\Sigma}_2$ is more concentrated (has a lower determinant) than $\hat{\Sigma}_1$. The condition $|\hat{\Sigma}_1| \neq 0$ in the C-step theorem is no real restriction because if $|\hat{\Sigma}_1| = 0$ the minimal objective value is already attained.

C-steps can be iterated until $|\hat{\Sigma}_{\text{new}}| = 0$ or $|\hat{\Sigma}_{\text{new}}| = |\hat{\Sigma}_{\text{old}}|$. The sequence of determinants obtained in this way must converge in a finite number of steps because there are only finitely many h-subsets. However, there is no guarantee that the final value $|\hat{\Sigma}_{\text{new}}|$ of the iteration process is the global minimum of the MCD objective function. Therefore, an approximate MCD solution can be obtained by taking many initial choices of H_1, applying C-steps to each and keeping the solution with lowest determinant.

To construct an initial subset H_1, a random $(p + 1)$-subset J is drawn and $\hat{\boldsymbol{\mu}}_0 := \text{ave}(J)$ and $\hat{\Sigma}_0 := C(J)$ are computed. (If $|\hat{\Sigma}_0| = 0$, then J can be extended by adding observations until $|\hat{\Sigma}_0| > 0$.) Then, for $i = 1, \ldots, n$ the distances $d_i^0 := d(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_0, \hat{\Sigma}_0)$ are computed and sorted. The initial H_1 subset then consists of the h observations with smallest distance d^0. This method yields better initial subsets than by drawing random h-subsets directly, because the probability of drawing an outlier-free subset is much higher when drawing $(p + 1)$-subsets than with h-subsets.

The FAST-MCD algorithm contains several computational improvements. Since each C-step involves the calculation of a covariance matrix, its determinant and the corresponding distances, using fewer C-steps considerably improves the speed of the algorithm. It turns out that after two C-steps, many runs that will lead to the global minimum already have a considerably smaller determinant. Therefore, the number of C-steps is reduced by applying only two C-steps to each initial subset and selecting the 10 different subsets with lowest determinants. Only for these 10 subsets further C-steps are taken until convergence.

This procedure is very fast for small sample sizes n, but when n grows the computation time increases due to the n distances that need to be calculated in each C-step. For large n FAST-MCD uses a partitioning of the data set, which avoids doing all the calculations on the entire data.

Note that the FAST-MCD algorithm is itself affine equivariant. Implementations of the FAST-MCD algorithm are available in R (as part of the packages *rrcov*, *robust* and *robustbase*), in S-PLUS (as the built-in function *cov.mcd*) and in SAS/IML Version 7 and SAS Version 9 (in *PROC ROBUSTREG*). There is also a Matlab version in LIBRA, a Matlab LIBrary for Robust Analysis (Verboven and Hubert 2005) which can be downloaded from wis.kuleuven.be/stat/robust. Moreover, it is available in the PLS toolbox of Eigenvector Research (www.eigenvector.com). Note that some functions use $\alpha = 0.5$ as default value, yielding a breakdown value of 50 %, whereas other implementations use $\alpha = 0.75$.

4.5 Other High-Breakdown Affine Equivariant Estimators

4.5.1 The Stahel–Donoho Estimator

The first affine equivariant estimator of location and scatter with 50 % breakdown value was the *Stahel–Donoho estimator* (Stahel 1981; Donoho 1982). It is constructed as follows. The Stahel–Donoho outlyingness of a univariate point x_i is given by

$$\mathrm{SDO}_i = \mathrm{SDO}^{(1)}(x_i, X_n) = \frac{|x_i - \mathrm{med}(X_n)|}{\mathrm{mad}(X_n)},$$

whereas the outlyingness of a multivariate $\mathbf{x}_i$ is defined as

$$\mathrm{SDO}_i = \mathrm{SDO}(\mathbf{x}_i, X_n) = \sup_{\mathbf{a}\in\mathbb{R}^p} \mathrm{SDO}^{(1)}\big(\mathbf{a}'\mathbf{x}_i, X_n\mathbf{a}\big). \tag{4.12}$$

The Stahel–Donoho estimator is then defined as a weighted mean and covariance matrix, where the weight function $W(t)$ is a strictly positive and nonincreasing function of SDO_i. If $t\,W(t)$ and $t^2W(t)$ are bounded, then the breakdown value is 50 %. Note that $\mathrm{mad}(X_n)$ in the denominator of SDO_i can be modified slightly to obtain the best possible finite-sample breakdown value (Gather and Hilker 1997). For more details on the Stahel–Donoho estimator see (Maronna and Yohai 1995).

To compute the Stahel–Donoho estimator, the number of directions $\mathbf{a}$ needs to be restricted to a finite set. These can be obtained by subsampling: take the directions orthogonal to hyperplanes spanned by random subsamples of size p.

In data sets with high contamination rate, it may happen that fewer than $p + 1$ observations obtain a weight $W(\mathrm{SDO}_i) > 0$ up to numerical precision, hence the Stahel–Donoho scatter matrix becomes singular. This can be remedied as in Hubert et al. (2005) and Debruyne and Hubert (2009) by replacing the smooth function W by weights that are set to 1 for the h points with lowest outlyingness, and to 0 for the others. This way enough data points are included to ensure nonsingularity, assuming the uncontaminated data were in general position.

4.5.2 The MVE Estimator

The *Minimum Volume Ellipsoid* (MVE) estimator (Rousseeuw 1985) of location is defined as the center of the ellipsoid with minimal volume which contains h observations with $[(n+p+1)/2] \leqslant h \leqslant n$. The corresponding scatter estimator is given by the matrix in the formula of this ellipsoid, multiplied by a consistency factor. This estimator has maximal breakdown value when $h = [(n+p+1)/2]$, but it is not asymptotically normal and more difficult to compute than the MCD estimator. A FAST-MVE algorithm is described in Maronna et al. (2006, p. 199).

4.5.3 S-Estimators

S-estimators of multivariate location and scatter (Rousseeuw and Leroy 1987; Lopuhaä 1989) are defined as the solution $(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ to the problem of

$$\text{minimizing } |\hat{\Sigma}| \quad \text{subject to} \quad \frac{1}{n}\sum_{i=1}^{n} \rho(d_i) = \delta$$

with $d_i = d(\mathbf{x}_i, \hat{\boldsymbol{\mu}}, \hat{\Sigma})$ as before and $0 < \delta < +\infty$.

The MVE estimator with 50 % breakdown value can be seen as a special case, obtained by taking $\rho(t) = I(|t| > \sqrt{\chi^2_{p,0.5}})$ and $\delta = 0.5$. However, this discontinuous ρ-function does not yield a good asymptotic behavior. Using a smooth ρ-function gives much better results. More precisely, the ρ function should satisfy:

- (R1) ρ is symmetric and twice continuously differentiable, with $\rho(0) = 0$;
- (R2) ρ is strictly increasing on an interval $[0, k]$ and constant on $[k, +\infty[$.

A standard choice is Tukey's bisquare function defined as

$$\rho_c(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & \text{if } |x| \leqslant c, \\ \frac{c^2}{6} & \text{if } |x| > c. \end{cases}$$

To obtain consistency at elliptical distributions, the constant δ is taken to be $E[\rho(\|Z\|)]$ with Z as in (4.3).

S-estimators are asymptotically normal (Davies 1987). Their efficiency at the normal model is somewhat better than the efficiency of the RMCD, especially in higher dimensions. For example, the diagonal element of the bisquare S scatter matrix with 50 % breakdown value has an asymptotic relative efficiency of 50.2 % for $p = 2$ and 92 % for $p = 10$. (Recall that the reweighted MCD achieves 45.5 % for $p = 2$ and 82 % for $p = 10$.) At the multivariate Student distribution the RMCD performs better in low dimensions (Croux and Haesbroeck 1999).

If X_n is in general position, the ratio $r = \delta/\rho(k)$ determines the breakdown value: if $r \leqslant (n-p)/2n$, then $\varepsilon^*(\hat{\boldsymbol{\mu}}, X_n) = \varepsilon^*(\hat{\Sigma}, X_n) = \lceil nr \rceil / n$. The maximal breakdown value (4.7) is attained when $r = (n-p)/2n$.

To obtain a bounded influence function, it is required that $\rho''(x)$ and $\rho'(x)/x$ are bounded and continuous (Lopuhaä 1989). The influence function of S-estimators can then be seen as a smoothed version of the MCD's influence function (Croux and Haesbroeck 1999).

S-estimators satisfy (4.9) and (4.10) of a non-monotone M-estimator, but they cannot be computed in this way because non-monotone M-estimators do not have a unique solution and there is no algorithm for them. To compute S-estimators, we can run the FAST-S algorithm which uses techniques similar to FAST-MCD and the FAST-S algorithm for regression (Salibian-Barrera and Yohai 2006). It is available in the R package `rrcov` as the function `CovSest`. A Matlab implementation is available from www.econ.kuleuven.be/public/NDBAE06/programs/.

4.5.4 MM-Estimators

Multivariate *MM-estimators* (Tatsuoka and Tyler 2000) are extensions of S-estimators. They are based on two loss functions ρ_0 and ρ_1 that satisfy (R1) and (R2). They are defined in two steps:

1. Let $(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma})$ be an S-estimator with loss function ρ_0. Denote $\hat{\sigma} = |\tilde{\Sigma}|^{1/2p}$.
2. The MM-estimator for location and shape $(\hat{\boldsymbol{\mu}}, \hat{\Gamma})$ minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1\Big(\big[(\mathbf{x}_i-\boldsymbol{\mu})'\Gamma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\big]^{1/2}/\hat{\sigma}\Big)$$

   among all $\boldsymbol{\mu}$ and all symmetric positive definite Γ with $|\Gamma| = 1$. The MM-estimator for the covariance matrix is then $\hat{\Sigma} = \hat{\sigma}^2\hat{\Gamma}$.

The idea is to estimate the scale by means of a very robust S-estimator, and then to estimate the location and shape using a different ρ function that yields a better efficiency. MM-estimators have the following properties:

- The location and shape estimates inherit the breakdown value of the initial scale: if $\rho_1(s) \leqslant \rho_0(s)$ for all $s > 0$ and $\rho_1(\infty) = \rho_0(\infty)$, then

$$\varepsilon_n^*(\hat{\boldsymbol{\mu}}, \hat{\Gamma}, \hat{\Sigma}; X_n) \geqslant \varepsilon_n^*(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}; X_n).$$

  Consequently, the values δ and k_0 in the initial S-estimator can be chosen to attain a certain breakdown value.
- The influence functions (and thus the asymptotic variance) of MM-estimators for location $\hat{\boldsymbol{\mu}}$ and shape $\hat{\Gamma}$ equal those of S-estimators with the function ρ_1. (Note that the influence function of the scatter estimator $\hat{\Sigma} = \hat{\sigma}^2\hat{\Gamma}$ is a mixture of the influence functions of the S-estimators with ρ_0 and ρ_1.) Hence, the constant k_1 in (R2) can be chosen to attain a certain efficiency for $\hat{\Gamma}$. In this way MM-estimators can have a higher efficiency than S-estimators, especially in dimensions $p < 15$ (Salibian-Barrera et al. 2006).

To compute MM-estimators, one first computes an S-estimator and then applies the iterative reweighting scheme described in Sect. 4.3. In R location-scatter MM-estimators are available as the function `CovMMest` in the package `rrcov`. This implementation uses a bisquare MM-estimator with 50 % breakdown value and 95 % efficiency at the normal model.

4.6 Robust Non Affine Equivariant Estimators

If one is willing to give up the affine equivariance requirement, certain robust location vectors and covariance matrices can be computed much faster.

4.6.1 Coordinatewise Median

The *coordinatewise median* is defined as

$$(\operatorname*{med}_i x_{i1}, \operatorname*{med}_i x_{i2}, \ldots, \operatorname*{med}_i x_{ip})'.$$

This estimator has a 50 % breakdown value and can be computed easily. But it is not affine equivariant, and it does not have to lie in the convex hull of the sample when $p \geqslant 3$. Consider for example the three unit vectors $(1, 0, 0)'$, $(0, 1, 0)'$ and $(0, 0, 1)'$ whose convex hull does not contain the coordinatewise median $(0, 0, 0)'$.

4.6.2 Spatial Median and Spatial Sign Covariance Matrix

The L^1 location estimator, also known as the *spatial median*, is defined as

$$\hat{\boldsymbol{\mu}} = \operatorname*{argmin}_{\boldsymbol{\mu}} \sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\mu}\|,$$

or equivalently as the $\boldsymbol{\mu}$ which satisfies

$$\sum_{i=1}^{n} \frac{\mathbf{x}_i - \boldsymbol{\mu}}{\|\mathbf{x}_i - \boldsymbol{\mu}\|} = 0. \tag{4.13}$$

It has a nice geometrical interpretation: take a point $\boldsymbol{\mu}$ in $\mathbb{R}^p$ and project all observations onto a sphere around $\boldsymbol{\mu}$. If the mean of these projections equals $\boldsymbol{\mu}$, then $\boldsymbol{\mu}$ is the spatial median.

The breakdown value of the L^1-median is 50 % and its influence function is bounded but it is not affine equivariant (Lopuhaä and Rousseeuw 1991). However, the L^1-median is *orthogonal equivariant*, i.e., it satisfies (4.4) with A any orthogonal matrix ($A' = A^{-1}$). This implies that the L^1-median transforms appropriately under all transformations that preserve Euclidean distances (such as translations, rotations, and reflections).

To compute the spatial median, note that (4.13) corresponds to (4.9) with $W_1(t) = 1/\sqrt{t}$. We can thus use the iterative algorithm with $\Sigma = I$. More refined algorithms are discussed in Fritz et al. (2012). For more on multivariate medians, see the contribution by Oja, Chap. 1.

The *spatial sign covariance matrix* (Visuri et al. 2000) is defined as the covariance matrix of the data points projected on the unit sphere around the spatial median:

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \frac{(\mathbf{x}_i - \hat{\boldsymbol{\mu}})}{\|\mathbf{x}_i - \hat{\boldsymbol{\mu}}\|} \frac{(\mathbf{x}_i - \hat{\boldsymbol{\mu}})'}{\|\mathbf{x}_i - \hat{\boldsymbol{\mu}}\|}. \tag{4.14}$$

4.6.3 The OGK Estimator

Maronna and Zamar (2002) presented a method to obtain positive definite and approximately affine equivariant robust scatter matrices starting from any pairwise robust scatter matrix. This method was applied to the robust covariance estimate of Gnanadesikan and Kettenring (1972). The resulting multivariate location and scatter estimates are called *orthogonalized Gnanadesikan–Kettenring* (OGK) estimates and are calculated as follows:

1. Let $m(\cdot)$ and $s(\cdot)$ be robust univariate estimators of location and scale.
2. Construct $\mathbf{z}_i = D^{-1}\mathbf{x}_i$ for $i = 1, \ldots, n$ with $D = \text{diag}(s(X_1), \ldots, s(X_p))$.
3. Compute the 'correlation matrix' U of the variables of $Z = (Z_1, \ldots, Z_p)$, given by $u_{jk} = 1/4(s(Z_j + Z_k)^2 - s(Z_j - Z_k)^2)$.
4. Compute the matrix E of eigenvectors of U and
   (a) project the data on these eigenvectors, i.e. $V = ZE$;
   (b) compute 'robust variances' of $V = (V_1, \ldots, V_p)$, i.e. $L = \text{diag}(s^2(V_1), \ldots, s^2(V_p))$;
   (c) Set the $p \times 1$ vector $\hat{\boldsymbol{\mu}}(Z) = E\boldsymbol{m}$ where $\boldsymbol{m} = (m(V_1), \ldots, m(V_p))'$, and compute the positive definite matrix $\hat{S}(Z) = ELE'$.
5. Transform back to X, i.e., $\hat{\boldsymbol{\mu}}_{\text{RAWOGK}} = D\hat{\mu}(Z)$ and $\hat{\Sigma}_{\text{RAWOGK}} = D\hat{S}(Z)D'$.

In the OGK algorithm $m(\cdot)$ is a weighted mean and $s(\cdot)$ is the τ-scale of Yohai and Zamar (1988). Step 2 makes the estimate scale equivariant, whereas the following steps are a kind of principal components that replace the eigenvalues of U (which may be negative) by robust variances. As in the FAST-MCD algorithm the estimate is improved by a weighting step, where the cutoff value in the weight function is now taken as $c = \chi^2_{p,0.9}\,\text{med}(d_1, \ldots, d_n)/\chi^2_{p,0.5}$ with $d_i = d(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_{\text{RAWOGK}}, \hat{\Sigma}_{\text{RAWOGK}})$.

4.6.4 Deterministic MCD Algorithm

As the FAST-MCD algorithm starts by drawing random subsets of size $p+1$, it does not necessarily give the same result at multiple runs of the algorithm. Moreover, it needs to draw many initial subsets in order to obtain at least one that is outlier-free. Recently a deterministic algorithm for robust location and scatter has been developed, called DetMCD (Hubert et al. 2012), which uses the same iteration steps as FAST-MCD but does not draw random subsets. Unlike the FAST-MCD it is permutation invariant, i.e. the result does not depend on the order of the observations in the data set. Furthermore, DetMCD runs even faster than FAST-MCD and is less sensitive to point contamination. Moreover, it is very close to affine equivariant.

DetMCD computes a small number of deterministic initial estimates, followed by concentration steps. First, each variable X_j is standardized by subtracting its median and dividing by the Q_n scale estimator of Rousseeuw and Croux (1993). This standardization makes the algorithm location and scale equivariant, i.e., (4.4) hold for any non-singular diagonal matrix $\mathbf{A}$. The standardized data set is denoted by the $n \times p$ matrix Z_n with rows $\mathbf{z}_i'$ $(i = 1, \ldots, n)$ and columns Z_j $(j = 1, \ldots, p)$.

Next, six preliminary estimates S_k are constructed $(k = 1, \ldots, 6)$ for the covariance or correlation of Z.

1. $S_1 = \operatorname{corr}(Y)$ with $Y_j = \tanh(Z_j)$ for $j = 1, \ldots, p$.
2. Let R_j be the ranks of the column Z_j, and put $S_2 = \operatorname{corr}(R)$. This is the Spearman correlation matrix of Z.
3. $S_3 = \operatorname{corr}(T)$ with $T_j = \Phi^{-1}((R_j - 1/3)/(n + 1/3))$.
4. The fourth scatter estimate is related to the spatial sign covariance matrix (4.14): define $\mathbf{k}_i = \mathbf{z}_i / \|\mathbf{z}_i\|$ for all i and let $S_4 = (1/n) \sum_{i=1}^{n} \mathbf{k}_i \mathbf{k}_i'$.
5. S_5 is the covariance matrix of the $\lceil n/2 \rceil$ standardized observations $\mathbf{z}_i$ with smallest norm, which corresponds to the first step of the BACON algorithm (Billor et al. 2000).
6. The sixth scatter estimate is the raw OGK estimator, where for $m(\cdot)$ and $s(\cdot)$ the median and Q_n are used.

As these S_k may have very inaccurate eigenvalues, the following steps are applied to each S_k. Note that the first two steps are similar to steps 4(a) and 4(b) of the OGK algorithm:

1. Compute the matrix E of eigenvectors of S_k and put $V = ZE$.
2. Estimate the scatter of Z by $\hat{\Sigma}_k(Z) = ELE'$ where $L = \operatorname{diag}(Q_n^2(V_1), \ldots, Q_n^2(V_p))$.
3. Estimate the center of Z by $\hat{\boldsymbol{\mu}}_k(Z) = \hat{\Sigma}_k^{1/2}(\operatorname{med}(Z \hat{\Sigma}_k^{-1/2}))$.

For all six estimates $(\hat{\boldsymbol{\mu}}_k(Z), \hat{\Sigma}_k(Z))$ the statistical distances $d_{ik} = d(\mathbf{z}_i, \hat{\boldsymbol{\mu}}_k(Z), \hat{\Sigma}_k(Z))$ are computed as in (4.2). For each initial estimate k, the $h_0 = \lfloor n/2 \rfloor$ observations with smallest d_{ik} are retained and the statistical distances (denoted as d_{ik}^*) based on these h_0 observations are computed. Then for all six estimates the h observations $\mathbf{x}_i$ with smallest d_{ik}^* are selected and C-steps are applied until convergence.

The solution with smallest determinant is called the raw DetMCD. Then a weighting step can be applied as in the FAST-MCD algorithm, yielding the final DetMCD.

Note that even though the OGK and DetMCD methods are not affine equivariant, it turns out that their deviation from affine equivariance is very small.

4.7 Conclusions

The assumption underlying this chapter is that the majority of the data can be modeled by an elliptical distribution, whereas there is no such restriction on any outliers that may occur. Unlike the classical mean and covariance matrix, robust estimators can withstand the effect of such outliers. Moreover, we saw in the example how the robust methods allow us to detect the outliers by means of their robust distances, which can for instance be visualized in a distance-distance plot like Fig. 4.4.

We advocate the use of robust estimators with a suitably high breakdown value, as these are the least affected by outliers. Our recommendation is to use a high-breakdown affine equivariant method such as MCD, S, or MM when the number of dimensions p is rather small, say up to 10. For higher dimensions, these methods become too computationally intensive, and then we recommend either OGK or DetMCD. The latter methods are close to affine equivariant, and can be computed faster and in higher dimensions.

References

Becker, C., & Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association*, *94*, 947–955.

Becker, C., & Gather, U. (2001). The largest nonidentifiable outlier: a comparison of multivariate simultaneous outlier identification rules. *Computational Statistics & Data Analysis*, *36*, 119–127.

Billor, N., Hadi, A., & Velleman, P. (2000). BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis*, *34*, 279–298.

Cator, E., & Lopuhaä, H. (2012). Central limit theorem and influence function for the MCD estimators at general multivariate distributions. *Bernoulli*, *18*, 520–551.

Croux, C., & Haesbroeck, G. (1999). Influence function and efficiency of the Minimum Covariance Determinant scatter matrix estimator. *Journal of Multivariate Analysis*, *71*, 161–190.

Davies, L. (1987). Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, *15*, 1269–1292.

Davies, P., & Gather, U. (2005). Breakdown and groups (with discussion and rejoinder). *The Annals of Statistics*, *33*, 977–1035.

Debruyne, M., & Hubert, M. (2009). The influence function of the Stahel–Donoho covariance estimator of smallest outlyingness. *Statistics & Probability Letters*, *79*, 275–282.

Donoho, D. (1982). *Breakdown properties of multivariate location estimators*. Ph.D. Thesis, Harvard University, Boston.

Fritz, H., Filzmoser, P., & Croux, C. (2012). A comparison of algorithms for the multivariate L1-median. *Computational Statistics*, *27*, 393–410.

Gather, U., & Hilker, T. (1997). A note on Tyler's modification of the MAD for the Stahel-Donoho estimator. *The Annals of Statistics*, *25*, 2024–2026.

Gnanadesikan, R., & Kettenring, J. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, *28*, 81–124.

Hampel, F., Ronchetti, E., Rousseeuw, P., & Stahel, W. (1986). *Robust statistics: the approach based on influence functions*. New York: Wiley.

Hubert, M., Rousseeuw, P., & Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal components analysis. *Technometrics*, *47*, 64–79.

Hubert, M., Rousseeuw, P., & Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, *21*, 618–637.

Lopuhaä, H. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *The Annals of Statistics*, *17*, 1662–1683.

Lopuhaä, H., & Rousseeuw, P. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, *19*, 229–248.

Maronna, R. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, *4*, 51–67.

Maronna, R., Martin, D., & Yohai, V. (2006). *Robust statistics: theory and methods*. New York: Wiley.

Maronna, R., & Yohai, V. (1995). The behavior of the Stahel–Donoho robust multivariate estimator. *Journal of the American Statistical Association*, *90*, 330–341.

Maronna, R., & Zamar, R. (2002). Robust estimates of location and dispersion for high-dimensional data sets. *Technometrics*, *44*, 307–317.

Pison, G., Van Aelst, S., & Willems, G. (2002). Small sample corrections for LTS and MCD. *Metrika*, *55*, 111–123.

Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*, 871–880.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, & W. Wertz (Eds.), *Mathematical statistics and applications* (Vol. B, pp. 283–297). Dordrecht: Reidel.

Rousseeuw, P., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*, 1273–1283.

Rousseeuw, P., & Leroy, A. (1987). *Robust regression and outlier detection*. New York: Wiley-Interscience.

Rousseeuw, P., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*, 212–223.

Rousseeuw, P., & van Zomeren, B. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, *85*, 633–651.

Salibian-Barrera, M., Van Aelst, S., & Willems, G. (2006). PCA based on multivariate MM-estimators with fast and robust bootstrap. *Journal of the American Statistical Association*, *101*, 1198–1211.

Salibian-Barrera, M., & Yohai, V. J. (2006). A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics*, *15*, 414–427.

Stahel, W. (1981). *Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*. Ph.D. Thesis, ETH Zürich.

Tatsuoka, K., & Tyler, D. (2000). On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. *The Annals of Statistics*, *28*, 1219–1243.

Verboven, S., & Hubert, M. (2005). LIBRA: a Matlab library for robust analysis. *Chemometrics and Intelligent Laboratory Systems*, *75*, 127–136.

Visuri, S., Koivunen, V., & Oja, H. (2000). Sign and rank covariance matrices. *Journal of Statistical Planning and Inference*, *91*, 557–575.

Yohai, V., & Zamar, R. (1988). High breakdown point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, *83*, 406–413.

Chapter 5
Upper and Lower Bounds for Breakdown Points

Christine H. Müller

5.1 Introduction

The breakdown point of an estimator introduced by Hampel (1971) is a simple and successful measure of the robustness of an estimator against changes of the observations. In particular, it is easy to understand the finite sample version of the breakdown point. Estimators with a high breakdown point are insensitive to a high amount of outlying observations. Moreover, they can be used to detect observations which do not follow the majority of the data. Some estimators have a breakdown of 50 % while in other situations the highest possible breakdown point is much smaller than 50 %. Therefore it is always important to know what is the highest possible breakdown point. Then it can be checked whether specific estimators can reach this upper bound. This can be done by deriving lower bounds for these estimators. Here general upper and lower bounds for the breakdown point are discussed.

Two finite sample breakdown point concepts are given in Sect. 5.2. In Sect. 5.3, a general upper bound is derived via the approach based on algebraic groups of transformations introduced by Davies and Gather (2005). While Davies and Gather (2005) develop this approach for the population version of the breakdown point, here this approach is used at once for the finite sample version of the breakdown point. This leads to a very simple characterization of the upper bound. Davies and Gather (2005) apply the approach to multivariate location and scatter estimation, univariate linear regression, logistic regression, the Michaelis–Menten model, and time series using different groups of transformations for each case. Regarding multivariate regression in Sect. 5.4, linear regression as well as multivariate location and scatter estimation can be treated here with the same approach. In particular the same group of transformations is used for the three cases. In Sect. 5.5, a general lower bound for the breakdown of some estimators based on the approach of d-fullness developed by Vandev (1993) is presented. With this approach, Müller and Neykov (2003)

C.H. Müller (✉)
Faculty of Statistics, TU Dortmund University, 44221 Dortmund, Germany
e-mail: cmueller@statistik.tu-dortmund.de

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_5, © Springer-Verlag Berlin Heidelberg 2013

derive lower bounds for generalized linear models like logistic regression and log-linear models and Müller and Schäfer (2010) obtain lower bounds for some non-linear models. This approach is used here in Sect. 5.6 to provide lower bounds for multivariate regression and simultaneous estimation of the scale and regression parameter in univariate regression. It is shown in particular that least trimmed squares estimators are attaining the upper bounds derived in Sect. 5.4.

5.2 Definitions of Finite Sample Breakdown Points

Let Θ be a finite dimensional parameter space, $\mathbf{z}_1, \dots, \mathbf{z}_N \in \mathcal{Z}$ a univariate or multivariate sample in $\mathcal{Z}$, and $\hat{\theta} : \mathcal{Z}^N \to \Theta$ an estimator for $\theta \in \Theta$. A general definition of the finite sample breakdown point of an estimator $\hat{\theta}$ at a sample $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)^\top$ is given by the minimum relative number of observations which must be corrupted so that the estimator breaks down. Breaking down means that the estimator can attain arbitrary values and in particular values arbitrarily close to the border of the parameter space.

To be more precise, let

$$\mathcal{Z}_M(\mathbf{Z}) := \left\{ \bar{\mathbf{Z}} \in \mathcal{Z}^N; \operatorname{card}\{n; \mathbf{z}_n \neq \bar{\mathbf{z}}_n\} \leq M \right\}$$

be the set of contaminated samples corrupted by at most M observations. If the estimates at samples of this set attain values arbitrarily close to the border of Θ, then we have a breakdown. If the parameter space Θ is for example $[0, \infty)$ as for scale estimation, then clearly the border of this set is given by 0 and ∞. If all estimates at samples of $\mathcal{Z}_M(\mathbf{Z})$ are included in a compact interval $[a, b] \subset (0, \infty)$, in particular $0 < a$, then these estimates do not become arbitrarily close to the border. Hence, there is no breakdown. Since $(0, \infty)$ is the interior of $[0, \infty)$, the property of "no breakdown" can be defined generally by the property that there exists a compact subset Θ_0 of the interior $\operatorname{int}(\Theta)$ of Θ so that all estimates at samples of $\mathcal{Z}_M(\mathbf{Z})$ are included in Θ_0. If such a compact set Θ_0 does not exist, then estimates at samples of $\mathcal{Z}_M(\mathbf{Z})$ reach the border so that we have breakdown. The smallest number M for which this happens provides then the breakdown point. One could define this M as breakdown point but it is better to use relative numbers, i.e., M divided by the sample size N. Hence, a general definition of the finite sample breakdown is as follows, see, e.g., Hampel et al. (1986, p. 97).

Definition 5.1 The breakdown point of an estimator $\hat{\theta} : \mathcal{Z}^N \to \Theta$ at $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)^\top \in \mathcal{Z}^N$ is defined as

$$\epsilon^*(\hat{\theta}, \mathbf{Z}) := \frac{1}{N} \min\big\{M;\ \text{there exists no compact set } \Theta_0 \subset \operatorname{int}(\Theta) \text{ with}$$
$$\big\{\hat{\theta}(\bar{\mathbf{Z}}); \bar{\mathbf{Z}} \in \mathcal{Z}_M(\mathbf{Z})\big\} \subset \Theta_0\big\}.$$

As soon as there exists a pseudometric d on Θ, then breakdown can be also defined by explosion, i.e., by the property that the distance $d(\hat{\theta}(\bar{\mathbf{Z}}), \hat{\theta}(\mathbf{Z}))$ between the estimates at the corrupted sample $\bar{\mathbf{Z}}$ and the original sample $\mathbf{Z}$ becomes arbitrarily large. This leads to the following definition, see, e.g., Donoho and Huber (1983), Davies and Gather (2005).

Definition 5.2

$$\epsilon^*(\hat{\theta}, \mathbf{Z}, d) := \frac{1}{N} \min\Big\{M; \sup_{\bar{\mathbf{Z}} \in \mathcal{Z}_M(\mathbf{Z})} d\big(\hat{\theta}(\bar{\mathbf{Z}}), \hat{\theta}(\mathbf{Z})\big) = \infty\Big\}.$$

If $\Theta = \mathbb{R}^p$, then the pseudometric can be chosen as the Euclidean metric $\|\cdot\|_p$. If $\Theta = [0, \infty) \subset \mathbb{R}$, for example for scale parameters, then an appropriate choice for the pseudometric is $d(\theta_1, \theta_2) = |\log(\theta_1 \cdot \theta_2^{-1})|$, see Davies and Gather (2005). This is again a metric but its extension to scatter matrices is only a pseudometric, as is discussed in Sect. 5.4.2. This pseudometric avoids the distinction between implosion breakdown point and explosion breakdown point as Rousseeuw and Hubert do, see Chap. 4.

Davies and Gather (2005) use the population version of the breakdown point and not the finite sample version of Definition 5.2. But they point out that the finite sample version is obtained by using the empirical distribution. They provide a general upper bound for the population version of Definition 5.2 using transformation groups on the sample space $\mathcal{Z}$. Here this approach is given at once in the sample version.

5.3 A General Upper Bound

A general upper bound for the finite sample breakdown point of Definition 5.2 can be given by the concept of equivariance. Equivariance is an important property of an estimator $\hat{\theta}$ if transformations of the data space are related to transformations of the parameter space. Then also the estimator should be transformed in the same way as the parameter is transformed. For example, a translation of multivariate observations $\mathbf{z}_1, \ldots, \mathbf{z}_N$ to $\mathbf{z}_1 + \gamma, \ldots, \mathbf{z}_N + \gamma$ is related to the translation of a location parameter by γ. Hence a location estimator should be also translated by γ. However, a scatter parameter is not influenced by a translation of the data and this should hold for a scatter estimator as well. Usually this property is distinguished as translation invariance. But here this property is included in the concept of equivariance since the parameter and the estimator varies in the same way. Generally, equivariance can be defined with respect to measurable transformations given by a group

$$\mathcal{G} := \{g; g : \mathcal{Z} \to \mathcal{Z}\}.$$

Recall that $\mathcal{G}$ is a group in algebraic sense with actions $\circ$ and unit element ι if and only if

- $g_1 \circ g_2 \in \mathcal{G}$ for all $g_1, g_2 \in \mathcal{G}$,
- $\iota \circ g = g = g \circ \iota$ for all $g \in \mathcal{G}$,
- for each $g \in \mathcal{G}$ there exists g^{-1} with $g \circ g^{-1} = \iota = g^{-1} \circ g$.

Definition 5.3 An estimator $\hat{\theta} : \mathcal{Z}^N \to \Theta$ is called equivariant with respect to a group $\mathcal{G}$ if there exists a group $\mathcal{H}_{\mathcal{G}} = \{h_g; g \in \mathcal{G}\}$ of transformations $h_g : \Theta \to \Theta$ such that for each $g \in \mathcal{G}$ there exists $h_g \in \mathcal{H}_{\mathcal{G}}$ with

$$\hat{\theta}\big(\big(g(\mathbf{z}_1), \dots, g(\mathbf{z}_N)\big)^\top\big) = h_g\big(\hat{\theta}\big((\mathbf{z}_1, \dots, \mathbf{z}_N)^\top\big)\big)$$

for all samples $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)^\top \in \mathcal{Z}^N$.

To derive the upper bound for the breakdown point, the following subset of $\mathcal{G}$ is needed

$$\mathcal{G}_1 := \Big\{g \in \mathcal{G}; \lim_{k \to \infty} \inf_{\theta \in \Theta} d\big(\theta, h_{g^k}(\theta)\big) = \infty\Big\}.$$

If $\mathcal{G}_1 = \emptyset$, then the group $\mathcal{G}$ is too small to produce transformed parameters arbitrarily far away from the original parameter. This transferres to the estimates by equivariance, since then the parameters and the estimator varies in the same way. Hence in this case, a breakdown in the sense of Definition 5.2 cannot be produced by transformations of the group $\mathcal{G}$. Therefore, $\mathcal{G}_1 \neq \emptyset$ is an important property for deriving an upper bound for the finite sample breakdown point.

Theorem 5.1 *If the estimator $\hat{\theta} : \mathcal{Z}^N \to \Theta$ is equivariant with respect to $\mathcal{G}$ and $\mathcal{G}_1 \neq \emptyset$, then*

$$\epsilon^*(\hat{\theta}, \mathbf{Z}, d) \leq \frac{1}{N} \left\lfloor \frac{N - \Delta(\mathbf{Z}) + 1}{2} \right\rfloor$$

for all $\mathbf{Z} \in \mathcal{Z}^N$, where

$$\Delta\big((\mathbf{z}_1, \dots, \mathbf{z}_N)^\top\big) := \max\big\{\text{card}\big\{n; g(\mathbf{z}_n) = \mathbf{z}_n\big\}; g \in \mathcal{G}_1\big\}$$

and $\lfloor x \rfloor$ is the largest integer m with $m \leq x$.

Note the more simple form of the quantity $\Delta(\mathbf{Z})$ compared with its form in the population version given by Davies and Gather (2005).

Proof Regard an arbitrary observation vector $\mathbf{Z}$. Let be $M = \lfloor \frac{N - \Delta(\mathbf{Z}) + 1}{2} \rfloor$ and $L = \Delta(\mathbf{Z})$. Then there exists $g \in \mathcal{G}_1$ so that without loss of generality $g(\mathbf{z}_n) = \mathbf{z}_n$ for $n = 1, \dots, L$. Then we also have $g^k(\mathbf{z}_n) = g^{k-1}(\mathbf{z}_n) = \dots = g^2(\mathbf{z}_n) = g \circ g(\mathbf{z}_n) = g(g(\mathbf{z}_n)) = g(\mathbf{z}_n) = \mathbf{z}_n$ for all $n = 1, \dots, L$ and all integer k. Define $\tilde{\mathbf{Z}}^k$ and $\bar{\mathbf{Z}}^k$ for any integer k by

$$\begin{aligned} \tilde{\mathbf{z}}_n^k &= \mathbf{z}_n \quad \text{for } n = 1, \dots, L \text{ and } L + M + 1, \dots, N, \\ \tilde{\mathbf{z}}_n^k &= g^k(\mathbf{z}_n) \quad \text{for } n = L + 1, \dots, L + M, \end{aligned}$$

and

$$\begin{aligned}\bar{\mathbf{z}}_n^k &= \mathbf{z}_n \quad \text{for } n = 1, \ldots, L+M,\\ \bar{\mathbf{z}}_n^k &= g^{-k}(\mathbf{z}_n) \quad \text{for } n = L+M+1, \ldots, N.\end{aligned}$$

Obviously, $\tilde{\mathbf{Z}}^k \in \mathcal{Z}_M(\mathbf{Z})$. Since $N-(L+M) = N-L-\lfloor\frac{N-L+1}{2}\rfloor \le N-L-\frac{N-L}{2} = \frac{N-L}{2} \le \lfloor\frac{N-L+1}{2}\rfloor$, we also have $\bar{\mathbf{Z}}^k \in \mathcal{Z}_M(\mathbf{Z})$. Moreover, it holds

$$\begin{aligned}&g^k\big(\bar{\mathbf{z}}_n^k\big) = g^k(\mathbf{z}_n) = \mathbf{z}_n = \tilde{\mathbf{z}}_n^k \quad \text{for } n = 1, \ldots, L,\\ &g^k\big(\bar{\mathbf{z}}_n^k\big) = g^k(\mathbf{z}_n) = \tilde{\mathbf{z}}_n^k \quad \text{for } n = L+1, \ldots, L+M,\\ &g^k\big(\bar{\mathbf{z}}_n^k\big) = g^k\big(g^{-k}(\mathbf{z}_n)\big) = g^k \circ g^{-k}(\mathbf{z}_n) = \mathbf{z}_n = \tilde{\mathbf{z}}_n^k \quad \text{for } n = L+M+1, \ldots, N.\end{aligned}$$

Since $g \in \mathcal{G}_1$ and $\hat{\theta}((g^k(\bar{\mathbf{z}}_1^k), \ldots, g^k(\bar{\mathbf{z}}_N^k))^\top) = h_{g^k}(\hat{\theta}((\bar{\mathbf{z}}_1^k, \ldots, \bar{\mathbf{z}}_N^k)^\top))$, we obtain

$$\begin{aligned}\lim_{k\to\infty} d\big(\hat{\theta}\big(\tilde{\mathbf{Z}}^k\big), \hat{\theta}\big(\bar{\mathbf{Z}}^k\big)\big) &= \lim_{k\to\infty} d\Big(\hat{\theta}\Big(\big(g^k\big(\bar{\mathbf{z}}_1^k\big), \ldots, g^k\big(\bar{\mathbf{z}}_N^k\big)\big)^\top\Big), \hat{\theta}\big(\bar{\mathbf{Z}}^k\big)\Big)\\ &= \lim_{k\to\infty} d\big(h_{g^k}\big(\hat{\theta}\big(\bar{\mathbf{Z}}^k\big)\big), \hat{\theta}\big(\bar{\mathbf{Z}}^k\big)\big) = \infty.\end{aligned}$$

Because of $d(\hat{\theta}(\tilde{\mathbf{Z}}^k), \hat{\theta}(\bar{\mathbf{Z}}^k)) \le d(\hat{\theta}(\tilde{\mathbf{Z}}^k), \hat{\theta}(\mathbf{Z})) + d(\hat{\theta}(\mathbf{Z}), \hat{\theta}(\bar{\mathbf{Z}}^k))$, at least one of $d(\hat{\theta}(\tilde{\mathbf{Z}}^k), \hat{\theta}(\mathbf{Z}))$ and $d(\hat{\theta}(\mathbf{Z}), \hat{\theta}(\bar{\mathbf{Z}}^k))$ must converge to ∞ for $k \to \infty$ as well. □

If $\Delta(\mathbf{Z}) = N$ then the upper bound for the finite breakdown point is 0. $\Delta(\mathbf{Z}) = N$ means that there exists a $g \in \mathcal{G}_1$ with $g(\mathbf{z}_n) = \mathbf{z}_n$ for all $n = 1, \ldots, N$. Then there are two possibilities for the estimate $\hat{\theta}(\mathbf{Z})$. One possibility is that $d(\hat{\theta}(\mathbf{Z}), h_{g^k}(\hat{\theta}(\mathbf{Z}))) < \infty$ for some k and $\lim_{k\to\infty} d(\hat{\theta}(\mathbf{Z}), h_{g^k}(\hat{\theta}(\mathbf{Z}))) = \infty$ which means that $\hat{\theta}(\mathbf{Z})$ is not unique since $h_{g^k}(\hat{\theta}(\mathbf{Z})) = \hat{\theta}((g^k(\mathbf{z}_1), \ldots, g^k(\mathbf{z}_N))) = \hat{\theta}(\mathbf{Z})$ is another estimate at $\mathbf{Z}$. For example, this is the case for regression parameters as shown below. The other possibility is that at once $d(\hat{\theta}(\mathbf{Z}), h_{g^k}(\hat{\theta}(\mathbf{Z}))) = \infty$ which means that $\hat{\theta}(\mathbf{Z})$ lies already at the border of the parameter space. For example, this is the case for scale and scatter parameters, see below.

5.4 Example: Multivariate Regression

The multivariate regression model is given by

$$\mathbf{y}_n^\top = \mathbf{x}_n^\top \mathbf{B} + \mathbf{e}_n^\top, \quad n = 1, \ldots, N, \tag{5.1}$$

where $\mathbf{y}_n \in \mathbb{R}^p$ is the observation vector, $\mathbf{x}_n \in \mathbb{R}^r$ the known regression vector, $\mathbf{B} \in \mathbb{R}^{r\times p}$ the unknown parameter matrix and $\mathbf{e}_n \in \mathbb{R}^p$ the error vector. Set $\mathbf{z} = (\mathbf{x}^\top, \mathbf{y}^\top)^\top \in \mathcal{Z} = \mathbb{R}^{r+p}$ and assume that $\mathbf{e}_1, \ldots, \mathbf{e}_N$ are realizations of i.i.d. random variables $E_1, \ldots, E_N$ with location parameter $\mathbf{0}_p$ and scatter matrix Σ, where $\mathbf{0}_p$ denotes the p-dimensional vector of zeros.

The interesting aspect of $\mathbf{B}$ shall be the linear aspect $\Lambda = \mathbf{LB}$ with $\mathbf{L} \in \mathbb{R}^{s\times r}$. Note that the whole matrix $\mathbf{B}$ is of interest if $\mathbf{L}$ is the $r \times r$-identity matrix. But to be more general, we consider also the case $\Lambda = \mathbf{LB}$ where $\mathbf{L}$ is not the identity matrix.

An example is the case where repeated measurements of patients getting two different drugs $D1$ and $D2$ are obtained at p days. Then we have $\mathbf{x}_n^\top = (1_{D1}(i_n), 1_{D1}(i_n))$, where i_n provides the drug of the n'th patient. The matrix $\mathbf{B} = (b_{ij})_{i=1,2,j=1,\dots,p} \in \mathbb{R}^{2\times p}$ contains the drug effects at the p days. An interesting aspect is then the row vector of differences $(b_{11} - b_{21}, \dots, b_{1p} - b_{2p})$ between the effects of drug $D1$ and $D2$ at the p days, i.e., $\Lambda = \mathbf{LB}$ with $\mathbf{L} = (1, -1)$.

Another application is forecasting as considered by Kharin, Chap. 14. If $\mathbf{x}_1 = \mathbf{v}(t_1), \dots, \mathbf{x}_N = \mathbf{v}(t_N)$ with $t_1 \le t_2 \le \dots \le t_N \in \mathbb{R}$ and $\mathbf{v} : \mathbb{R} \to \mathbb{R}^r$ is a known regression function, then $\Lambda = \mathbf{v}(\tau)^\top \mathbf{B}$ is a forecast for the expected value of an observation at $\tau > t_N$.

We consider here the problem of estimating Λ in Sect. 5.4.1 and of estimating Σ in Sect. 5.4.2. In both cases, we can use the following group of transformations

$$\mathcal{G} = \left\{ g_{\mathbf{A},\mathbf{B}} : \mathcal{Z} \to \mathcal{Z};\ \mathbf{A} \in \mathbb{R}^{p\times p} \text{ is regular}, \mathbf{B} \in \mathbb{R}^{r\times p} \right\}$$

with $g_{\mathbf{A},\mathbf{B}}((\mathbf{x}^\top, \mathbf{y}^\top)^\top) = (\mathbf{x}^\top, \mathbf{y}^\top \mathbf{A} + \mathbf{x}^\top \mathbf{B})^\top$. The unit element of this group is $\iota = g_{\mathbf{I}_p, \mathbf{0}_{r\times p}}$, where $\mathbf{0}_{r\times p}$ is the $r \times p$-dimensional zero matrix and $\mathbf{I}_p$ the p-dimensional identity matrix. The inverse of $g_{\mathbf{A},\mathbf{B}}$ is given by $g_{\mathbf{A}^{-1}, -\mathbf{B}\mathbf{A}^{-1}}$.

5.4.1 Estimation of a Linear Aspect of the Regression Parameters

Transforming $\mathbf{y}_n^\top = \mathbf{x}_n^\top \mathbf{B}_0 + \mathbf{e}_n^\top$ to $\tilde{\mathbf{y}}_n = \mathbf{y}_n^\top \mathbf{A} + \mathbf{x}_n^\top \mathbf{B}$ leads to

$$\tilde{\mathbf{y}}_n = \mathbf{x}_n^\top \mathbf{B}_0 \mathbf{A} + \mathbf{e}_n^\top \mathbf{A} + \mathbf{x}_n^\top \mathbf{B} = \mathbf{x}_n^\top (\mathbf{B}_0 \mathbf{A} + \mathbf{B}) + \mathbf{e}_n^\top \mathbf{A} \tag{5.2}$$

so that $\Lambda = \mathbf{L}\mathbf{B}_0$ becomes $\tilde{\Lambda} = \mathbf{L}(\mathbf{B}_0 \mathbf{A} + \mathbf{B}) = \Lambda \mathbf{A} + \mathbf{LB}$. Hence, an estimator $\hat{\theta} = \hat{\Lambda} : \mathcal{Z}^N \to \mathbb{R}^{s\times p}$ for $\theta = \Lambda = \mathbf{LB} \in \mathbb{R}^{s\times p}$ should satisfy

$$\hat{\mathbf{\Lambda}}\left(\left(g_{\mathbf{A},\mathbf{B}}(\mathbf{z}_1), \dots, g_{\mathbf{A},\mathbf{B}}(\mathbf{z}_n) \right)^\top \right) = h_{g_{\mathbf{A},\mathbf{B}}}\left(\hat{\Lambda}(\mathbf{Z}) \right)$$

with $h_{g_{\mathbf{A},\mathbf{B}}}(\Lambda) = \Lambda \mathbf{A} + \mathbf{LB}$ for all $g_{\mathbf{A},\mathbf{B}} \in \mathcal{G}$, i.e., it should be scatter equivariant and translation equivariant. With $\mathcal{G}$, also

$$\mathcal{H}_{\mathcal{G}} = \left\{ h_{g_{\mathbf{A},\mathbf{B}}} : \mathbb{R}^{s\times p} \to \mathbb{R}^{s\times p};\ \mathbf{A} \in \mathbb{R}^{p\times p} \text{ is regular}, \mathbf{B} \in \mathbb{R}^{r\times p} \right\}$$

is a group of transformations.

If $\mathbf{LB} = \mathbf{0}_{s\times p}$, then $\Lambda = \mathbf{0}_{s\times p}$ satisfies

$$d\left(\Lambda, h_{g^n_{\mathbf{A},\mathbf{B}}}(\Lambda)\right) = d\left(\mathbf{0}_{s\times p}, \mathbf{0}_{s\times p}\mathbf{A}^n\right) = d(\mathbf{0}_{s\times p}, \mathbf{0}_{s\times p}) = 0$$

for any pseudometric d on $\mathbb{R}^{s\times p}$. Hence $\mathbf{LB} \neq \mathbf{0}_{s\times p}$ is necessary for $g_{\mathbf{A},\mathbf{B}} \in \mathcal{G}_1$. Moreover, we have $\Lambda = h_{g_{\mathbf{A},\mathbf{B}}}(\Lambda) = \Lambda \mathbf{A} + \mathbf{LB}$ if and only if $\mathbf{LB} = \Lambda(\mathbf{I}_p - \mathbf{A})$ so that

$$\mathcal{G}_1 = \left\{ g_{\mathbf{A},\mathbf{B}} \in \mathcal{G};\ \mathbf{LB} \neq \mathbf{0}_{s\times p} \text{ and } \mathbf{LB} \neq \Lambda(\mathbf{I}_p - \mathbf{A}) \text{ for all } \Lambda \in \mathbb{R}^{s\times p} \right\}.$$

Set $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^\top$. Now we are going to show that $\Delta(\mathbf{Z})$ is the maximum number of regressors $\mathbf{x}_n$ so that the univariate linear aspect $\mathbf{L}\beta$ with $\beta \in \mathbb{R}^r$ is not identifiable at these regressors, i.e. $\Delta(\mathbf{Z})$ is the nonidentifiability parameter $\mathcal{N}_\lambda(\mathbf{X})$ defined in Müller (1995) for univariate regression, see also Müller (1997).

Definition 5.4 $\mathbf{L}\beta$ is identifiable at $D = \{\mathbf{x}_{n_1}, \ldots, \mathbf{x}_{n_I}\}$ if for all $\beta \in \mathbb{R}^r$

$$\mathbf{x}_{n_i}^\top \beta = 0 \quad \text{for } i = 1, \ldots, I \quad \text{implies } \mathbf{L}\beta = 0.$$

If $\mathbf{X}_D = (\mathbf{x}_{n_1}, \ldots, \mathbf{x}_{n_I})^\top$, then it is well known that $\mathbf{L}\beta$ is identifiable at $D = \{\mathbf{x}_{n_1}, \ldots, \mathbf{x}_{n_I}\}$ if and only if $\mathbf{L} = \mathbf{K}\mathbf{X}_D$ for some $\mathbf{K} \in \mathbb{R}^{s \times I}$, see, e.g., Müller (1997, p. 6).

Definition 5.5 The nonidentifiability parameter $\mathcal{N}_\lambda(\mathbf{X})$ for estimating $\lambda = \mathbf{L}\beta$ in univariate regression, i.e., $\beta \in \mathbb{R}^r$, is defined as

$$\begin{aligned}\mathcal{N}_\lambda(\mathbf{X}) &:= \max\big\{\text{card}\big\{n; \mathbf{x}_n^\top \beta = 0\big\}; \beta \in \mathbb{R}^r \text{ with } \lambda = \mathbf{L}\beta \neq 0\big\} \\ &= \max\{\text{card}\, D; \lambda = \mathbf{L}\beta \text{ is not identifiable at } D\}.\end{aligned}$$

Theorem 5.2 *For estimating the linear aspect* $\mathbf{LB}$ *of the regression parameter* $\mathbf{B}$ *in the regression model* (5.1), *we have*

$$\Delta(\mathbf{Z}) = \mathcal{N}_\lambda(\mathbf{X}).$$

Proof Let be $g_{\mathbf{A},\mathbf{B}} \in \mathcal{G}_1$ and assume that there exists $\mathbf{z}_{n_1}, \ldots, \mathbf{z}_{n_I}$ with $g_{\mathbf{A},\mathbf{B}}(\mathbf{z}_{n_i}) = \mathbf{z}_{n_i}$ for $i = 1, \ldots, I$.

If $\mathbf{A} = \mathbf{I}_p$, then it holds $g_{\mathbf{A},\mathbf{B}}(\mathbf{z}) = \mathbf{z} = (\mathbf{x}^\top, \mathbf{y}^\top)^\top$ if and only if $\mathbf{x}^\top \mathbf{B} = \mathbf{0}_{1\times p}$ so that $\Delta(\mathbf{Z}) \geq \max\{\text{card}\{n; \mathbf{x}_n^\top \beta = 0\}; \beta \in \mathbb{R}^p \text{ with } \mathbf{L}\beta \neq 0\}$ since $\mathbf{LB} \neq \mathbf{0}_{s\times p}$ for $g_{\mathbf{A},\mathbf{B}} \in \mathcal{G}_1$. In this case, $\mathbf{LB} \neq \Lambda(\mathbf{I}_{p\times p} - \mathbf{A})$ is always satisfied for all $\Lambda \in \mathbb{R}^{s\times p}$ so that it is no restriction.

Now consider $\mathbf{A} \neq \mathbf{I}_p$. Assume that $\mathbf{L}\beta$ is identifiable at $D = \{\mathbf{x}_{n_1}, \ldots, \mathbf{x}_{n_I}\}$ with $I = \Delta(\mathbf{Z})$. Then there exists $\mathbf{K} \in \mathbb{R}^{s\times I}$ such that $\mathbf{L} = \mathbf{K}\mathbf{X}_D$. Set $\mathbf{Y}_D = (\mathbf{y}_{n_1}, \ldots, \mathbf{y}_{n_I})^\top$. Since $g_{\mathbf{A},\mathbf{B}}(\mathbf{z}_{n_i}) = \mathbf{z}_{n_i}$ if and only if $\mathbf{x}_{n_i}^\top \mathbf{B} = \mathbf{y}_{n_i}^\top (\mathbf{I}_p - \mathbf{A})$, we obtain the contradiction

$$\mathbf{LB} = \mathbf{K}\mathbf{X}_D\mathbf{B} = \mathbf{K}\begin{pmatrix}\mathbf{x}_{n_1}^\top \mathbf{B} \\ \vdots \\ \mathbf{x}_{n_I}^\top \mathbf{B}\end{pmatrix} = \mathbf{K}\begin{pmatrix}\mathbf{y}_{n_1}^\top (\mathbf{I}_p - \mathbf{A}) \\ \vdots \\ \mathbf{y}_{n_I}^\top (\mathbf{I}_p - \mathbf{A})\end{pmatrix} = \mathbf{K}\mathbf{Y}_D(\mathbf{I}_p - \mathbf{A})$$

since $g_{\mathbf{A},\mathbf{B}} \in \mathcal{G}_1$ implies $\mathbf{LB} \neq \Lambda(\mathbf{I}_p - \mathbf{A})$ for all $\Lambda \in \mathbb{R}^{s\times p}$. This means that $\mathbf{L}\beta$ cannot be identifiable at $D = \{\mathbf{x}_{n_1}, \ldots, \mathbf{x}_{n_I}\}$ so that $\Delta(\mathbf{Z}) = I \leq \max\{\text{card}\{n; \mathbf{x}_n^\top \beta = 0\}; \beta \in \mathbb{R}^p \text{ with } \mathbf{L}\beta \neq 0\}$. □

From the proof of Theorem 5.2, it is clear that the assertion of Theorem 5.2 holds also without using the scatter equivariance of the estimator $\hat{\Lambda}$. See also Sects. 5.4.1.1 and 5.4.1.2.

5.4.1.1 Location Model

A special case of multivariate regression is multivariate location with $\mathbf{x}_n = 1$ for all $n = 1, \ldots, N$, where $\mathbf{B} \in \mathbb{R}^{1\times p}$ is the parameter of interest. In this case, identifiability holds always so that $\Delta(\mathbf{Z}) = 0$. Hence, the upper bound of the finite sample breakdown point is $\frac{1}{N}\lfloor\frac{N+1}{2}\rfloor$ which is the highest possible upper bound. This result was obtained by Davies and Gather (2005) using only the translation group

$$\mathcal{G}^L = \left\{g_{\mathbf{I}_p,\mathbf{B}} : \mathcal{Z} \to \mathcal{Z}; \mathbf{B} \in \mathbb{R}^{1\times p}\right\}$$

so that

$$\mathcal{G}_1^L = \left\{g_{\mathbf{I}_p,\mathbf{B}} \in \mathcal{G}^L; \mathbf{B} \neq \mathbf{0}_{1\times p}\right\}.$$

They wrote in their rejoinder that $\mathcal{G}_1^L$ would be empty if scatter transformations are considered as well. But $\mathcal{G}_1^L$ becomes larger if a larger group of transformation is regarded.

For the special case of univariate data, i.e., $p = 1$, with location parameter $\mu \in \mathbb{R}$, the condition

$$\mathbf{0}_{s\times p} \neq \mathbf{LB} \neq \Lambda(\mathbf{I}_p - \mathbf{A}) \quad \text{for all } \Lambda \in \mathbb{R}^{1\times p} \tag{5.3}$$

becomes

$$0 \neq b \neq \mu(1-a) \quad \text{for all } \mu \in \mathbb{R}, \tag{5.4}$$

where $a, b \in \mathbb{R}$ replace $\mathbf{A}$ and $\mathbf{B}$. Since condition (5.4) is only satisfied for $a = 1$ we have

$$\mathcal{G}_1^L = \mathcal{G}_1$$

so that indeed it does not matter if the scatter (here scale) equivariance is additionally demanded.

For univariate data, the upper bound $\frac{1}{N}\lfloor\frac{N+1}{2}\rfloor$ is attained by the median. Multivariate extensions of the median, which are scatter and translation equivariant, are Tukey's half space median and the Oja median. But the Oja median has only a finite sample breakdown point of $\frac{1}{N}$, see the contribution by Oja (Chap. 1), and the finite sample breakdown point of Tukey's half space median lies between $\frac{1}{p+1}$ and $\frac{1}{3}$, see Donoho and Gasko (1992). Another scatter and translation equivariant estimator is the location estimator given by the minimum covariance determinant (MCD) estimator. It has a finite sample breakdown point of $\frac{1}{N}\lfloor\frac{N-p+1}{2}\rfloor$, see the contribution by Rousseeuw and Hubert, Chap. 4. As far as the author knows, there is no scatter and translation equivariant location estimator which attains the upper bound $\frac{1}{N}\lfloor\frac{N+1}{2}\rfloor$ for $p > 1$.

5.4.1.2 Univariate Regression

Another special case of multivariate regression is univariate regression with $p = 1$, where the unknown parameter $\mathbf{B}$ is $\beta \in \mathbb{R}^r$. The result

$$\Delta(\mathbf{Z}) = \mathcal{N}_\lambda(\mathbf{X})$$

is obtained by Müller (1995) using only the transformations $g_\mathbf{b}((\mathbf{x}^\top, y)^\top) = (\mathbf{x}^\top, y + \mathbf{x}^\top\mathbf{b})^\top$ in a proof similar to that of Theorem 5.1, see also Müller (1997).

The special case $\mathbf{L}\beta = \beta$ is considered by Davies (1993) who derives the upper bound for the population version of the breakdown point. Davies and Gather (2005) provide this result as an example of the approach via groups.

Using the translation group

$$\mathcal{G}^R = \left\{ g_{\mathbf{b}} : \mathcal{Z} \to \mathcal{Z};\, \mathbf{b} \in \mathbb{R}^r \right\}$$

with $g_{\mathbf{b}}((\mathbf{x}^\top, y)^\top) = (\mathbf{x}^\top, y + \mathbf{x}^\top \mathbf{b})^\top$, as Davies and Gather (2005) propose, leads to

$$\mathcal{G}_1^R = \left\{ g_{\mathbf{b}} \in \mathcal{G}^R;\, \mathbf{b} \neq \mathbf{0}_r \right\}.$$

But since condition (5.3) becomes here

$$\mathbf{0}_r \neq \mathbf{b} \neq \beta(1-a) \quad \text{for all } \beta \in \mathbb{R}^r,$$

with $\mathbf{b} \in \mathbb{R}^r$ and $a \in \mathbb{R}$, it is again only satisfied for $a = 1$ so that

$$\mathcal{G}_1^R = \mathcal{G}_1.$$

Hence as for location estimation, the restriction to translations is no real restriction here.

5.4.2 Scatter Estimation

The transformation (5.2) of the regression model (5.1) leads to an error term of the form $\mathbf{e}_n^\top \mathbf{A}$. If $\mathbf{e}_n$ is a realization of a random variable E_n with scatter matrix $\mathbf{\Sigma}$, then $\mathbf{A}^\top \mathbf{e}_n$ is a realization of a random variable $\mathbf{A}^\top E_n$ with scatter matrix $\mathbf{A}^\top \mathbf{\Sigma} \mathbf{A}$. Hence, an estimator $\hat{\theta} = \hat{\mathbf{\Sigma}} : \mathcal{Z}^N \to \mathcal{S}$ of the scatter matrix $\Sigma \in \mathcal{S} = \{\mathbf{A} \in \mathbb{R}^{p \times p};\, \mathbf{A}$ is symmetric and positive definite$\}$, should satisfy

$$\hat{\mathbf{\Sigma}}\left(\left(g_{\mathbf{A},\mathbf{B}}(\mathbf{z}_1), \ldots, g_{\mathbf{A},\mathbf{B}}(\mathbf{z}_n) \right)^\top \right) = h_{g_{\mathbf{A},\mathbf{B}}}\left(\hat{\Sigma}(\mathbf{Z}) \right)$$

with $h_{g_{\mathbf{A},\mathbf{B}}}(\Sigma) = \mathbf{A}^\top \mathbf{\Sigma} \mathbf{A}$ for all $g_{\mathbf{A},\mathbf{B}} \in \mathcal{G}$, i.e., it should be scatter equivariant and translation invariant. With $\mathcal{G}$, also

$$\mathcal{H}_{\mathcal{G}} = \left\{ h_{g_{\mathbf{A},\mathbf{B}}} = h_{\mathbf{A}} : \mathcal{S} \to \mathcal{S};\, \mathbf{A} \in \mathbb{R}^{p \times p} \text{ is regular} \right\}$$

is a group of transformations. An appropriate pseudometric on $\mathcal{S}$ is given by

$$d(\Sigma_1, \Sigma_2) := \left| \log\left(\det\left(\Sigma_1 \Sigma_2^{-1} \right) \right) \right|.$$

It holds $d(\Sigma_1, \Sigma_2) = 0$ if and only if $\det(\Sigma_1 \Sigma_2^{-1}) = 1$. This is not only satisfied by $\Sigma_1 = \Sigma_2$, since e.g. diagonal matrices like $\operatorname{diag}(1, 1)$ and $\operatorname{diag}(\frac{1}{2}, 2)$ are satisfying this as well. Hence d is not a metric. But it is a pseudometric because it is always greater than 0 and it satisfies the triangle inequality. Since $\det(\mathbf{A}^\top \Sigma_1 \mathbf{A} \Sigma_2^{-1}) = \det(\Sigma_1 \Sigma_2^{-1})$ as soon as $\det(\mathbf{A}) = 1$, $\mathcal{G}_1$ is given by

$$\mathcal{G}_1 = \left\{ g_{\mathbf{A},\mathbf{B}} \in \mathcal{G};\, \det(\mathbf{A}) \neq 1 \right\}.$$

Since $g_{\mathbf{A},\mathbf{B}}(\mathbf{z}) = \mathbf{z}$ if and only if $\mathbf{x}^\top \mathbf{B} = \mathbf{y}^\top (\mathbf{I}_p - \mathbf{A})$, we have at once the following theorem.

Theorem 5.3 *For estimating the scatter matrix Σ in the regression model* (5.1), *we have*

$$\Delta(\mathbf{Z}) = \max\big\{\mathrm{card}\big\{n; \mathbf{x}_n^\top \mathbf{B} = \mathbf{y}_n^\top (\mathbf{I}_p - \mathbf{A})\big\};\ \mathbf{B} \in \mathbb{R}^{r\times p},\\ \mathbf{A} \in \mathbb{R}^{p\times p} \textit{ is regular with } \det(\mathbf{A}) \neq 1\big\}.$$

5.4.2.1 Location Model

In the special case of multivariate location with $\mathbf{x}_n = 1$ for all $n = 1, \ldots, N$ and $\mathbf{B} \in \mathbb{R}^{1\times p}$, it holds $g_{\mathbf{A},\mathbf{B}}(\mathbf{z}) = \mathbf{z}$ if and only if $\mathbf{B} = \mathbf{y}^\top(\mathbf{I}_p - \mathbf{A})$. Hence $\{\mathbf{y} \in \mathbb{R}^p;\ \mathbf{B} = \mathbf{y}^\top(\mathbf{I}_p - \mathbf{A})\}$ is a hyperplane in $\mathbb{R}^p$. Conversely, if $\{\mathbf{y} \in \mathbb{R}^p;\ \mathbf{c}^\top = \mathbf{y}^\top\mathbf{C}\}$ is an arbitrary hyperplane in $\mathbb{R}^p$, then it can be assumed that $\det(\mathbf{I}_p - \mathbf{C}) \neq 1$ so that $g_{\mathbf{I}_p-\mathbf{C},\mathbf{c}^\top} \in \mathcal{G}_1$. This implies that $\Delta(\mathbf{Z})$ is the maximum number of observations lying in a hyperplane. According to Theorem 5.1, the upper bound of the breakdown point of an equivariant scatter estimator is given by the maximum number of observations in a hyperplane. If all observations are lying in a hyperplane, then the estimated scatter matrix is not of full rank, i.e. at the border of the parameter space, so that the finite sample breakdown point is 0. If only a subset of observations are lying in a hyperplane, then the majority of the remaining observations determines the estimation of the scatter matrix by any reasonable estimator. Hence, corruption of this majority can lead to a breakdown so that the upper bound for the finite sample breakdown point is $\frac{1}{N}\lfloor\frac{N-\Delta(\mathbf{Z})+1}{2}\rfloor$.

This upper bound attains its highest value for observations in general position. Per definition, observations $\mathbf{y}_1, \ldots, \mathbf{y}_N \in \mathbb{R}^p$ are in general position if any subset of $p+1$ observations are not lying in a hyperplane. But since p points are lying in the hyperplane of $\mathbb{R}^p$ spanned by these points, an upper bound for the breakdown point is always $\frac{1}{N}\lfloor\frac{N-p+1}{2}\rfloor$, see also the contribution by Rousseeuw and Hubert, Chap. 4. The population version of this result was originally given by Davies (1993) and derived by group equivariance in Davies and Gather (2005). The upper bound $\frac{1}{N}\lfloor\frac{N-p+1}{2}\rfloor$ is for example attained by the minimum covariance determinant (MCD) estimator, see the contribution by Rousseeuw and Hubert.

For the one-dimensional case ($p = 1$), the upper bound of the breakdown point of a scale equivariant and translation invariant scale estimator is determined by the maximum number of repeated observations. Note that we have here $\mathbf{A} = a \in \mathbb{R}$ with $a \neq 1$, $\mathbf{B} = b \in \mathbb{R}$ so that $g_{\mathbf{A},\mathbf{B}}(\mathbf{z}_n) = g_{a,b}(\mathbf{z}_n) = \mathbf{z}_n$ if and only if $b = y_n(1-a)$ or equivalently $y_n = \frac{b}{1-a}$. Hence, $\Delta(\mathbf{Z}) = \max\{\mathrm{card}\{n;\ y_n = c\};\ c \in \mathbb{R}\}$. Here the highest value of the upper bound is given by pairwise different observations. This highest upper bound is for example attained by the median absolute deviation (MAD). However, it can happen that the upper bound is not attained by the median absolute deviation if observations are repeated. Davies and Gather (2007) give the following example

$$1.0,\ 1.8,\ 1.3,\ 1.3,\ 1.9,\ 1.1,\ 1.3,\ 1.6,\ 1.7,\ 1.3,\ 1.3.$$

The median absolute deviation of this sample is 0.2. But as soon as one observation unequal to 1.3 is replaced by 1.3, the median absolute deviation is 0. Hence the breakdown point of this sample is $\frac{1}{11}$. However, since 1.3 is repeated five times, the upper bound for the breakdown point is

$$\frac{1}{11}\left\lfloor \frac{11-5+1}{2} \right\rfloor = \frac{3}{11}.$$

5.4.2.2 Univariate Regression

In the special case of univariate regression with $p = 1$, i.e., $\Sigma = \sigma^2 \in \mathbb{R}^+$, the condition

$$\mathbf{x}^\top \mathbf{B} = \mathbf{y}^\top (\mathbf{I}_p - \mathbf{A})$$

becomes

$$\mathbf{x}^\top \beta = y(1-a) \quad \Longleftrightarrow \quad y = \mathbf{x}^\top \tilde{\beta}$$

with $\beta \in \mathbb{R}^r$, $1 \neq a \in \mathbb{R}$ and $\tilde{\beta} = \frac{1}{1-a}\beta$. This means that $\Delta(\mathbf{Z})$ is the maximum number $\mathcal{E}(\mathbf{X})$ of observations satisfying an exact fit. Thereby observations $y_1, \ldots, y_N$ are satisfying an exact fit if there exists $\beta \in \mathbb{R}^r$ so that $y_n = \mathbf{x}_n^\top \beta$ for all $n = 1, \ldots, N$.

Definition 5.6 The exact fit parameter is defined as

$$\mathcal{E}(\mathbf{X}) := \max\big\{\text{card}\big\{n;\, y_n = \mathbf{x}_n^\top \beta\big\};\, \beta \in \mathbb{R}^p\big\}.$$

Hence, we have here

$$\Delta(\mathbf{Z}) = \mathcal{E}(\mathbf{X}).$$

Clearly, if all observations are satisfying an exact fit, i.e., $\mathcal{E}(\mathbf{X}) = N$, then the variance σ^2 should be estimated by 0 which provides a finite sample breakdown point of 0. Again, if only a subset of the observations satisfy an exact fit, then the majority of the remaining data determines completely the behaviour of an equivariant scale estimator and can cause breakdown.

5.5 A General Lower Bound for Some Estimators

Since there are always estimators with a breakdown point of $\frac{1}{N}$ or even 0, a lower bound can be only valid for some special estimators. He we consider estimators of the form

$$\hat{\theta}(\mathbf{Z}) := \arg\min_{\theta \in \Theta} s(\mathbf{Z}, \theta)$$

with $s : \mathcal{Z}^N \times \Theta \to \mathbb{R}$, where $s(\mathbf{Z}, \theta)$ can be bounded from below and above by some quality functions $q : \mathcal{Z} \times \Theta \to \mathbb{R}$. These quality functions can be residuals but also some negative loglikelihood functions as considered in Müller and Neykov

(2003). Set $q_n(\mathbf{Z}, \theta) = q(\mathbf{z}_n, \theta)$ for $n = 1, \ldots, N$ and $q_{(1)}(\mathbf{Z}, \theta) \leq \cdots \leq q_{(N)}(\mathbf{Z}, \theta)$. Then there shall exists $\alpha, \beta \in \mathbb{R}$ with $\alpha \neq 0$ and $h \in \{1, \ldots, N\}$ such that

$$\alpha q_{(h)}(\mathbf{Z}, \theta) \leq s(\mathbf{Z}, \theta) \leq \beta q_{(h)}(\mathbf{Z}, \theta) \tag{5.5}$$

for all $\mathbf{Z} \in \mathcal{Z}^N$ and $\theta \in \Theta$. In particular, h-trimmed estimators given by

$$\hat{\theta}_h(\mathbf{Z}) := \arg\min_{\theta \in \Theta} \sum_{n=1}^{h} q_{(n)}(\mathbf{Z}, \theta)$$

are satisfying condition (5.5). In particular least trimmed squares (LTS) estimators, where $q_n(\mathbf{Z}, \theta)$ is the squared residuum, are of this form. But also S-estimators are satisfying condition (5.5), see, e.g., Rousseeuw and Leroy (2003, pp. 135–139).

For deriving a lower bound for the breakdown point, Definition 5.1 for the breakdown point is used. This definition is checking whether the estimators are remaining in a compact subset of the parameter space. Via compact sets, Vandev (1993) develops the concept of d-fullness which is used by Vandev and Neykov (1998) to estimate this breakdown point for trimmed estimators. A modification of this concept, used in Müller and Neykov (2003), bases on the following definitions.

Definition 5.7 A function $\gamma : \Theta \to \mathbb{R}$ is called sub-compact if the set $\{\theta \in \Theta;\ \gamma(\theta) \leq c\}$ is contained in a compact set $\Theta_c \subset \text{int}(\Theta)$ for all $c \in \mathbb{R}$.

Definition 5.8 A finite set $\Gamma = \{\gamma_n : \Theta \to \mathbb{R};\ n = 1, \ldots, N\}$ of functions is called d-full if for each $\{n_1, \ldots, n_d\} \subset \{1, \ldots, N\}$ the function γ given by $\gamma(\theta) := \max\{\gamma_{n_k}(\theta);\ k = 1, \ldots, d\}$ is sub-compact.

For example, consider a quadratic regression model with $\mathbf{x}_n = \mathbf{v}(t_n) = (1, t_n, t_n^2)^\top \in \mathbb{R}^3$, $t_n \in [-1, 1]$, and $\beta = (\beta_0, \beta_1, \beta_2)^\top \in \mathbb{R}^3$ and let $q(\mathbf{z}_n, \beta) = (y_n - \mathbf{x}_n^\top \beta)^2$ be the quality function. If $N = 8$, $t_1 = t_2 = t_3 = -1$, $t_4 = t_5 = t_6 = 0$, $t_7 = t_8 = 1$, then $\{q(\mathbf{z}_n, \cdot);\ n = 1, \ldots, 8\}$ is not 6-full, since $\gamma(\beta) = \max\{q(\mathbf{z}_n, \beta);\ n = 1, \ldots, 6\}$ is not sub-compact. To see that $\gamma(\beta)$ is not sub-compact, consider c_0 with $\sqrt{c_0} = \max\{|y_n|;\ n = 1, \ldots, 8\}$. Then $-\sqrt{c_0} \leq y_n \leq \sqrt{c_0}$ for all y_n imply

$$\begin{aligned}
&\left\{\beta \in \mathbb{R}^3;\ \gamma(\beta) \leq c\right\} \\
&\quad = \left\{\beta \in \mathbb{R}^3;\ \left(y_n - \mathbf{x}_n^\top \beta\right)^2 \leq c_0 \text{ for } n = 1, \ldots, 6\right\} \\
&\quad = \left\{\beta \in \mathbb{R}^3;\ -\sqrt{c_0} \leq y_n - \mathbf{x}_n^\top \beta \leq \sqrt{c_0} \text{ for } n = 1, \ldots, 6\right\} \\
&\quad = \left\{\beta \in \mathbb{R}^3;\ y_n - \sqrt{c_0} \leq -\mathbf{x}_n^\top \beta \leq y_n + \sqrt{c_0} \text{ for } n = 1, \ldots, 6\right\} \\
&\quad \supset \left\{\beta \in \mathbb{R}^3;\ 0 \leq -\mathbf{x}_n^\top \beta \leq 0 \text{ for } n = 1, \ldots, 6\right\} \\
&\quad = \left\{\beta \in \mathbb{R}^3;\ \beta_0 - \beta_1 + \beta_2 = 0 \text{ and } \beta_0 = 0\right\} = \left\{\beta \in \mathbb{R}^3;\ \beta_1 = \beta_2\right\},
\end{aligned}$$

which is a hyperplane in $\mathbb{R}^3$ and thus not a compact subset in $\text{int}(\mathbb{R}^3) = \mathbb{R}^3$. However, $\{q(\mathbf{z}_n, \cdot);\ n = 1, \ldots, 8\}$ is 7-full, since any subset of seven observations contains the experimental conditions $t_n = -1$, $t_n = 0$, and $t_n = 1$, see also Lemma 5.1 below.

Theorem 5.4 (Müller and Neykov 2003) *If the estimator $\hat{\theta}$ satisfies condition* (5.5) *and $\{q_n(\mathbf{Z},\cdot); n=1,\dots,N\}$ is d-full, then*

$$\epsilon^*(\hat{\theta},\mathbf{Z}) \geq \frac{1}{N}\min\{N-h+1, h-d+1\}.$$

The lower bound of Theorem 5.4 is maximized if the trimming factor h satisfies $\lfloor\frac{N+d}{2}\rfloor \leq h \leq \lfloor\frac{N+d+1}{2}\rfloor$. A simple consequence of this fact is the following result concerning trimmed estimators.

Theorem 5.5 *Assume that $\{q_n(\mathbf{Z},\cdot); n=1,\dots,N\}$ is d-full and $\lfloor\frac{N+d}{2}\rfloor \leq h \leq \lfloor\frac{N+d+1}{2}\rfloor$. Then the breakdown point of any trimmed estimator $\hat{\theta}_h$ satisfies*

$$\epsilon^*(\hat{\theta}_h,\mathbf{Z}) \geq \frac{1}{N}\left\lfloor\frac{N-d+2}{2}\right\rfloor.$$

5.6 Example: Regression

5.6.1 Multivariate Regression

Consider again multivariate regression with $x \in \mathbb{R}^r$, $y \in \mathbb{R}^p$ and unknown matrix $\mathbf{B} \in \mathbb{R}^{r\times p}$ of regression parameters. An appropriate quality function for estimating $\mathbf{B}$ is given by

$$q(\mathbf{z},\theta) = q(\mathbf{x},\mathbf{y},\mathbf{B}) = \left\|\mathbf{y}-\mathbf{B}^\top\mathbf{x}\right\|_p^2 = \left(\mathbf{y}^\top - \mathbf{x}^\top\mathbf{B}\right)\left(\mathbf{y}-\mathbf{B}^\top\mathbf{x}\right). \tag{5.6}$$

The h-trimmed estimator $\hat{\mathbf{B}}$ for $\mathbf{B}$ can be determined by calculating the least squares estimator

$$\hat{\mathbf{B}}_I(\mathbf{Y}) = \left(\mathbf{X}_I^\top\mathbf{X}_I\right)^{-1}\mathbf{X}_I^\top\mathbf{Y}_I$$

for each subsample $I=\{n_1,\dots,n_h\} \subset \{1,\dots,N\}$ for which the inverse of $\mathbf{X}_I^\top\mathbf{X}_I$ exists, where $\mathbf{X}_I = (\mathbf{x}_{n_1},\dots,\mathbf{x}_{n_h})^\top$ and $\mathbf{Y}_I = (\mathbf{y}_{n_1},\dots,\mathbf{y}_{n_h})^\top$. Then $\hat{\mathbf{B}}(\mathbf{Y})$ is that $\hat{\mathbf{B}}_{I_*}(\mathbf{Y}_{I_*})$ with

$$I_* = \arg\min\left\{\sum_{j=1}^{h}\left\|\mathbf{y}_{n_j} - \hat{\mathbf{B}}_I(\mathbf{Y}_I)^\top\mathbf{x}_{n_j}\right\|_p^2; I=\{n_1,\dots,n_h\}\subset\{1,\dots,N\}\right\}.$$

However, exact computation is only possible for small sample sizes N. For larger sample sizes, a genetic algorithm with concentration step like that proposed by Neykov and Müller (2003) can be used, see also Rousseeuw and Driessen (2006).

Note that the inverse of $\mathbf{X}_I^\top\mathbf{X}_I$ always exists as soon as h is larger than the non-identifiability parameter $\mathcal{N}_\beta(\mathbf{X})$ with $\lambda=\beta$. The subset estimator $\hat{\mathbf{B}}_I$ is scatter and translation equivariant so that $\hat{\mathbf{B}}_{I_*}$ is translation equivariant. However $\hat{\mathbf{B}}_{I_*}$ is only scatter (scale) equivariant if $p = 1$. Otherwise it is only scatter equivariant with respect to orthogonal matrices $\mathbf{A}$ since then

$$\begin{aligned}
&q\big(\mathbf{x}, \mathbf{A}^\top \mathbf{y} + \mathbf{B}^\top \mathbf{x}, \hat{\mathbf{B}}_I(\mathbf{YA} + \mathbf{XB})\big)\\
&= \big\|\mathbf{A}^\top \mathbf{y} + \mathbf{B}^\top \mathbf{x} - \hat{\mathbf{B}}_I(\mathbf{YA} + \mathbf{XB})^\top \mathbf{x}\big\|_p\\
&= \big\|\mathbf{A}^\top \mathbf{y} + \mathbf{B}^\top \mathbf{x} - \big(\mathbf{A}^\top \mathbf{Y}_I^\top + \mathbf{B}^\top \mathbf{X}_I^\top\big)\mathbf{X}_I\big(\mathbf{X}_I^\top \mathbf{X}_I\big)^{-1}\mathbf{x}\big\|_p\\
&= \big\|\mathbf{A}^\top \mathbf{y} - \mathbf{A}^\top \mathbf{Y}_I^\top \mathbf{X}_I\big(\mathbf{X}_I^\top \mathbf{X}_I\big)^{-1}\mathbf{x}\big\|_p = \big\|\mathbf{A}^\top \mathbf{y} - \mathbf{A}^\top \hat{\mathbf{B}}_I(\mathbf{Y})^\top \mathbf{x}\big\|_p\\
&= \big(\mathbf{y}^\top - \mathbf{x}^\top \hat{\mathbf{B}}_I(\mathbf{Y})\big)\mathbf{AA}^\top\big(\mathbf{y} - \hat{\mathbf{B}}_I(\mathbf{Y})^\top \mathbf{x}\big) = q\big(\mathbf{x}, \mathbf{y}, \hat{\mathbf{B}}_I(\mathbf{Y})\big)
\end{aligned}$$

for all $I = \{n_1, \ldots, n_h\} \subset \{1, \ldots, N\}$.

The d-fullness is given here by the nonidentifiability parameter $\mathcal{N}_\beta(\mathbf{X})$. This is an extension of the result in Müller and Neykov (2003) where it is proved for univariate generalized linear models.

Lemma 5.1 *If the quality function q is given by* (5.6), *then* $\{q_n(\mathbf{Z}, \cdot); n = 1, \ldots, N\}$ *is d-full with* $d = \mathcal{N}_\beta(\mathbf{X}) + 1$.

Proof Consider any $I \subset \{1, \ldots, N\}$ with cardinality $\mathcal{N}_\beta(\mathbf{X}) + 1$. Then the triangle inequality provides for any $c \in \mathbb{R}$

$$\begin{aligned}
&\Big\{\mathbf{B} \in \mathbb{R}^{r\times p}; \max_{i\in I} q_i(\mathbf{z}_i, \mathbf{B}) \le c\Big\}\\
&= \Big\{\mathbf{B} \in \mathbb{R}^{r\times p}; \max_{i\in I} \big\|\mathbf{y}_i - \mathbf{B}^\top \mathbf{x}_i\big\|_p \le \sqrt{c}\Big\}\\
&\subset \Big\{\mathbf{B} \in \mathbb{R}^{r\times p}; \max_{i\in I} \big\|\mathbf{B}^\top \mathbf{x}_i\big\|_p - \|\mathbf{y}_i\|_p \le \sqrt{c}\Big\}\\
&\subset \Big\{\mathbf{B} \in \mathbb{R}^{r\times p}; \max_{i\in I} \big\|\mathbf{B}^\top \mathbf{x}_i\big\|_p \le \sqrt{c} + \max_{i\in I} \|\mathbf{y}_i\|_p\Big\}\\
&= \Big\{\mathbf{B} \in \mathbb{R}^{r\times p}; \max_{i\in I} \big\|\mathbf{B}^\top \mathbf{x}_i\big\|_p \le \sqrt{\tilde{c}}\Big\}\\
&= \left\{(\mathbf{b}_1, \ldots, \mathbf{b}_p) \in \mathbb{R}^{r\times p}; \max_{i\in I} \sum_{j=1}^p \big(\mathbf{b}_j^\top \mathbf{x}_i\big)^2 \le \tilde{c}\right\}\\
&\subset \left\{(\mathbf{b}_1, \ldots, \mathbf{b}_p) \in \mathbb{R}^{r\times p}; \frac{1}{\mathcal{N}_\beta(\mathbf{X}) + 1} \sum_{i\in I} \sum_{j=1}^p \mathbf{b}_j^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{b}_j \le \tilde{c}\right\}\\
&= \left\{(\mathbf{b}_1, \ldots, \mathbf{b}_p) \in \mathbb{R}^{r\times p}; \frac{1}{\mathcal{N}_\beta(\mathbf{X}) + 1} \sum_{j=1}^p \mathbf{b}_j^\top \sum_{i\in I} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{b}_j \le \tilde{c}\right\}.
\end{aligned}$$

The definition of $\mathcal{N}_\beta(\mathbf{X})$ implies that the matrix $\sum_{i\in I} \mathbf{x}_i \mathbf{x}_i^\top$ is of full rank. Hence the set $\{(\mathbf{b}_1, \ldots, \mathbf{b}_p) \in \mathbb{R}^{r\times p}; \frac{1}{\mathcal{N}_\beta(\mathbf{X})+1} \sum_{j=1}^p \mathbf{b}_j^\top \sum_{i\in I} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{b}_j \le \tilde{c}\}$ is bounded and therefore included in a compact subset of $\mathbb{R}^{r\times p}$. □

Since the upper bound for the breakdown point given by Theorems 5.1 and 5.2 holds also for estimators which are not scatter equivariant, the combination of these theorems, Theorem 5.5 and Lemma 5.1 provides the following result. This result is derived for univariate regression already in Müller (1995).

Theorem 5.6 *If* $\lfloor\frac{N+\mathcal{N}_\beta(\mathbf{X})+1}{2}\rfloor \le h \le \lfloor\frac{N+\mathcal{N}_\beta(\mathbf{X})+2}{2}\rfloor$, *then the breakdown point of the trimmed estimator* $\hat{\mathbf{B}}_h$ *for* $\mathbf{B}$ *with quality function given by* (5.6) *satisfies*

$$\epsilon^*(\hat{\mathbf{B}}_h, \mathbf{Z}) = \frac{1}{N}\left\lfloor\frac{N-\mathcal{N}_\beta(\mathbf{X})+1}{2}\right\rfloor.$$

Müller (1995) shows Theorem 5.6 not only for estimating β but also for general linear aspects $\lambda = \mathbf{L}\beta$ of univariate regression models. Thereby $\mathcal{N}_\beta(\mathbf{X})$ must be only replaced by $\mathcal{N}_\lambda(\mathbf{X})$ in Theorem 5.6. However in this case the lower bound cannot be derived via d-fullness. In Müller (1995), the lower bound is proved directly for trimmed estimators, see also Müller (1997). This proof holds also for multivariate regression so that Theorem 5.6 holds also for linear aspects $\Lambda = \mathbf{LB}$ of multivariate regression.

5.6.2 Univariate Regression with Simultaneous Scale Estimation

If simultaneously the regression parameter $\beta \in \mathbb{R}^r$ and the scale parameter $\sigma \in \mathbb{R}^+$ in a univariate regression model shall be estimated, then the following quality function can be used

$$q(\mathbf{z}, \beta, \sigma) = q(\mathbf{x}, y, \beta, \sigma) = \frac{1}{2}\left(\frac{y-\mathbf{x}^\top\beta}{\sigma}\right)^2 + \log(\sigma). \tag{5.7}$$

In Müller and Neykov (2003), a slightly more general quality function is considered. But for simplicity, the quality function (5.7) shall be used here. The h-trimmed estimator $(\hat{\beta}, \hat{\sigma})$ for (β, σ) can be determined by calculating the maximum likelihood estimators

$$\hat{\beta}_I(\mathbf{y}) = \left(\mathbf{X}_I^\top \mathbf{X}_I\right)^{-1}\mathbf{X}_I^\top \mathbf{y}_I$$

and

$$\hat{\sigma}_I(\mathbf{y}) = \sqrt{\frac{1}{h}\sum_{j=1}^{h}\left(y_{n_j} - \mathbf{x}_{n_j}^\top \hat{\beta}_I(\mathbf{y})\right)^2}$$

for each subsample $I = \{n_1, \dots, n_h\} \subset \{1, \dots, N\}$, where $\mathbf{y}_I = (\mathbf{y}_{n_1}, \dots, y_{n_h})^\top$ and again $\mathbf{X}_I = (\mathbf{x}_{n_1}, \dots, \mathbf{x}_{n_h})^\top$. Then $(\hat{\beta}(\mathbf{y}), \hat{\sigma}(\mathbf{y}))$ is that $(\hat{\beta}_{I_*}(\mathbf{y}), \hat{\sigma}_{I_*}(\mathbf{y}))$ with

$$I_* = \arg\min\left\{\sum_{j=1}^{h} q\left(\mathbf{x}_{n_j}, y_{n_j}, \hat{\beta}_I(\mathbf{y}), \hat{\sigma}_I(\mathbf{y})\right);\ I = \{n_1, \dots, n_h\} \subset \{1, \dots, N\}\right\}.$$

$\hat{\beta}_I$ is translation equivariant and scale equivariant and $\hat{\sigma}_I$ is translation invariant and scale equivariant. Therefore, we have

$$\begin{aligned}
&q\big(\mathbf{x}, ya + \mathbf{x}^\top\beta, \hat{\beta}_I(\mathbf{y}a + \mathbf{X}\beta), \hat{\sigma}_I(\mathbf{y}a + \mathbf{X}\beta)\big)\\
&= \frac{1}{2}\left(\frac{ya + \mathbf{x}^\top\beta - \mathbf{x}^\top(\hat{\beta}_I(\mathbf{y})a + \beta)}{\hat{\sigma}_I(\mathbf{y})a}\right)^2 + \log\big(\hat{\sigma}_I(\mathbf{y})a\big)\\
&= q\big(\mathbf{x}, y, \hat{\beta}_I(\mathbf{y}), \hat{\sigma}_I(\mathbf{y})\big) + \log(a)
\end{aligned}$$

for all $I = \{n_1, \ldots, n_h\} \subset \{1, \ldots, N\}$ so that $\hat{\beta}_{I_*}$ is translation equivariant and scale equivariant and $\hat{\sigma}_{I_*}$ is translation invariant and scale equivariant.

Since the simultaneous estimator $(\hat{\beta}, \hat{\sigma})$ for (β, σ) breaks down when one of its components breaks down, an upper bound of the breakdown point of $(\hat{\beta}, \hat{\sigma})$ is $\frac{1}{N}\lfloor\frac{N-\max\{\mathcal{N}_\beta(\mathbf{X}), \mathcal{E}(\mathbf{X})\}+1}{2}\rfloor$ according to Sects. 5.4.1 and 5.4.2.

Deriving a lower bound for the breakdown point, Müller and Neykov (2003) implicitly assume that the exact fit parameter $\mathcal{E}(\mathbf{X})$ is zero. Here we extend this result for the case that it does not necessarily need to be zero.

Theorem 5.7 *If the quality function q is given by (5.7), then $\{q_n(\mathbf{Z}, \cdot); n = 1, \ldots, N\}$ is d-full with $d = \max\{\mathcal{N}_\beta(\mathbf{X}), \mathcal{E}(\mathbf{X})\} + 1$.*

Proof We have to show that γ given by

$$\gamma(\beta, \sigma) := \max_{i \in I} \frac{1}{2}\left(\frac{y_i - \mathbf{x}_i^\top\beta}{\sigma}\right)^2 + \log(\sigma)$$

is sub-compact for all $I \subset \{1, \ldots, N\}$ with cardinality $\max\{\mathcal{N}_\beta(\mathbf{X}), \mathcal{E}(\mathbf{X})\} + 1$. Take any $c \in \mathbb{R}$ and set $\tilde{\beta}(\sigma) := \arg\min\{\gamma(\beta, \sigma); \beta \in \mathbb{R}^r\}$ and $\tilde{\sigma}(\beta) := \arg\min\{\gamma(\beta, \sigma); \sigma \in \mathbb{R}^+\}$. Then $\tilde{\beta}(\sigma)$ is independent of σ such that $\tilde{\beta}(\sigma) =: \tilde{\beta}$. Setting

$$\gamma_1(\sigma) := \gamma\big(\tilde{\beta}(\sigma), \sigma\big) = \max_{i \in I} \frac{1}{2}\left(\frac{y_i - \mathbf{x}_i^\top\tilde{\beta}}{\sigma}\right)^2 + \log(\sigma)$$

we see that γ_1 is a sub-compact function since I has cardinality greater than $\mathcal{E}(\mathbf{X})$. Hence, there exists a compact set $\Theta_1 \subset \operatorname{int}(\mathbb{R}^+)$ such that $\{\sigma; \gamma_1(\sigma) \le c\} \subset \Theta_1$. Moreover, we have that with $\eta(\beta) := \max_{i \in I} |y_i - \mathbf{x}_i^\top\beta|$

$$\tilde{\sigma}(\beta) = \eta(\beta)$$

so that

$$\gamma_2(\beta) := \gamma\big(\beta, \tilde{\sigma}(\beta)\big) = \frac{1}{2} + \log\big(\eta(\beta)\big).$$

The proof of Lemma 5.1 provides that η is sub-compact. Since the logarithm is monotone also γ_2 is sub-compact so that $\{\beta; \gamma_2(\beta) \le c\} \subset \Theta_2$ for some compact set $\Theta_2 \subset \operatorname{int}(\mathbb{R}^r)$. Then we have

$$\begin{aligned}&\left\{(\beta,\sigma)\in\mathbb{R}^r\times\mathbb{R}^+;\gamma(\beta,\sigma)\le c\right\}\\&\quad\subset\left\{(\beta,\sigma)\in\mathbb{R}^r\times\mathbb{R}^+;\gamma_1(\sigma)\le c\text{ and }\gamma_2(\beta)\le c\right\}\subset\Theta_2\times\Theta_1\end{aligned}$$

so that γ is sub-compact. □

Theorem 5.8 *If* $\lfloor\frac{N+\max\{\mathcal{N}_\beta(\mathbf{X}),\mathcal{E}(\mathbf{X})\}+1}{2}\rfloor\le h\le\lfloor\frac{N+\max\{\mathcal{N}_\beta(\mathbf{X}),\mathcal{E}(\mathbf{X})\}+2}{2}\rfloor$, *then the breakdown point of the trimmed estimator* $(\hat\beta,\hat\sigma)_h$ *for* (β,σ) *with quality function given by* (5.7) *satisfies*

$$\epsilon^*\big((\hat\beta,\hat\sigma)_h,\mathbf{Z}\big)=\frac{1}{N}\left\lfloor\frac{N-\max\{\mathcal{N}_\beta(\mathbf{X}),\mathcal{E}(\mathbf{X})\}+1}{2}\right\rfloor.$$

5.7 Conclusions

The concept of equivariance provides a general upper bound for the finite sample breakdown point. This leads in particular to upper bounds for estimators for the location parameter, the regression parameter and the scatter matrix in multivariate location and regression models. These upper bounds can be reached by specific estimators for univariate models, for estimating the scatter matrix in multivariate location models and for estimating the regression parameters in multivariate regression models. In particular, the finite sample breakdown point of trimmed estimators are reaching this upper bound. This can be seen by a general lower bound for the finite sample breakdown given by the concept of d-fullness. This approach can be also applied to generalized linear models or nonlinear models where equivariance properties cannot be used. However, up to now scatter estimation cannot be treated with this. Another problem is correlation where the parameter space is $[-1, 1]$. Here up to now, the concept of equivariance was not successfully applied to get upper bounds for the finite sample breakdown point. The same holds for related problems like principal component analysis, discriminant analysis and applications on compositional data as considered by Filzmoser and Hron, see Chap. 8.

References

Davies, P. L. (1993). Aspects of robust linear regression. *The Annals of Statistics*, *21*, 1843–1899.

Davies, P. L., & Gather, U. (2005). Breakdown and groups (with discussion). *The Annals of Statistics*, *33*, 977–1035.

Davies, P. L., & Gather, U. (2007). The breakdown point—examples and counterexamples. *REVSTAT Statistical Journal*, *5*, 1–17.

Donoho, D. L., & Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, *20*, 1803–1827.

Donoho, D. L., & Huber, P. J. (1983). The notion of breakdown point. In P. J. Bickel, K. Doksum, & J. L. Hodges (Eds.), *A Festschrift for Erich L. Lehmann* (pp. 157–184). Belmont: Wadsworth.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, *42*, 1887–1896.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics. The approach based on influence functions*. New York: Wiley.

Müller, Ch. H. (1995). Breakdown points for designed experiments. *Journal of Statistical Planning and Inference*, *45*, 413–427.

Müller, Ch. H. (1997). *Lecture notes in statistics: Vol. 124. Robust planning and analysis of experiments*. New York: Springer.

Müller, Ch. H., & Neykov, N. (2003). Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and Inference*, *116*, 503–519.

Müller, Ch. H., & Schäfer, Ch. (2010). Designs with high breakdown point in nonlinear models. In A. Giovagnoli, A. C. Atkinson, & B. Torsney (Eds.), *mODa 9—advances in model-oriented design and analysis* (pp. 137–144). Heidelberg: Physica-Verlag.

Neykov, N., & Müller, Ch. H. (2003). Breakdown point and computation of trimmed likelihood estimators in generalized linear models. In R. Dutter, P. Filzmoser, U. Gather, & P. J. Rousseeuw (Eds.), *Developments in robust statistics* (pp. 277–286). Heidelberg: Physica-Verlag.

Rousseeuw, P. J., & Leroy, A. M. (2003). *Robust regression and outlier detection*. New York: Wiley.

Rousseeuw, P. J., & Driessen, K. (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, *12*, 29–45.

Vandev, D. L. (1993). A note on breakdown point of the least median squares and least trimmed squares. *Statistics & Probability Letters*, *16*, 117–119.

Vandev, D., & Neykov, N. (1998). About regression estimators with high breakdown point. *Statistics*, *32*, 111–129.

Chapter 6
The Concept of α-Outliers in Structured Data Situations

Sonja Kuhnt and André Rehage

6.1 Introduction

In every statistical data analysis, somehow surprising observations can occur which deviate strongly from the remaining observations or the assumed model. On the one hand, these observations may contain important pieces of information about the data-generating process. On the other hand, they might simply be measurement or reporting errors. Regardless of which origin the observation has, it is commonly named "outlier". There are numerous ways to detect outliers, with no strategy outperforming others in every situation. Besides non-parametric procedures, e.g., based on depth measures, also model-based strategies exist.

In order to be able to detect outliers, it first needs to be specified what is meant by an outlier. In this contribution, we discuss the notion of α-outliers as introduced by Davies and Gather (1993). The basic idea is that there exists a pattern which is supported by the majority of the data. Observations which are strongly deviating from this pattern are understood as outliers. Within the α-outlier concept, the pattern is the statistical model one has in mind for the data generating mechanism. Observations which lie in a region with low probability and are thereby surprising are understood as outliers. The general idea of α-outliers can be applied to basically any statistical model. The so-called outlier region usually is uniquely defined for a given statistical distribution. However, within the analysis of observed data sets this is often only specified up to some unknown parameters of the assumed class of distributions, resulting in the necessity of outlier identification procedures.

This chapter is structured as follows: Sect. 6.2 reviews the general definition of α-outlier regions. One-step approaches towards the detection of α-outliers in a

S. Kuhnt (✉) · A. Rehage
Faculty of Statistics, TU Dortmund University, 44221 Dortmund, Germany
e-mail: kuhnt@statistik.tu-dortmund.de

A. Rehage
e-mail: rehage@statistik.tu-dortmund.de

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_6, © Springer-Verlag Berlin Heidelberg 2013

data set are discussed in Sect. 6.3. In the remainder of this chapter, we focus on three specific situations of structured data situations: regression models in Sect. 6.4, contingency tables in Sect. 6.5 and graphical models in Sect. 6.6.

6.2 The Concept of α-Outliers

Besides numerous informal ways to narrow the term "outlier", there are also mathematical approaches to the concept of outliers and their detection (Barnett and Lewis 1994). We focus on so-called α-outliers, a concept which can be applied to basically any data situation where we have a statistical model in mind.

The notion of α-outlier regions as the largest $\alpha \times 100\ \%$ most improbable region of the target distribution goes back to Davies and Gather (1993). Starting from the univariate normal situation the treatment of α-outliers soon extends to the multivariate normal case (Becker and Gather 1999, 2001), to exponential samples (Schultze and Pawlitschko 2002) as well as more structured data situations (Wellmann and Gather 2003; Gather et al. 2002; Boscher 1992; Kuhnt 2004; Kuhnt and Pawlitschko 2005). Gather et al. (2003) generalize the original definition to arbitrary families of distributions.

Definition 6.1 (Gather et al. 2003) Let $\mathcal{P}$ be a family of distributions on a measurable space $(\mathcal{X}, \mathcal{A})$ which is dominated by a σ-finite measure ν such that $P \in \mathcal{P}$ has ν-density f. For $P \in \mathcal{P}$ let $\operatorname{supp}(P)$ denote the support of P and set $\operatorname{supp}(\mathcal{P}) = \bigcup_{P \in \mathcal{P}} \operatorname{supp}(P)$. For a given $\alpha \in (0, 1)$ the *α-outlier region* of $P \in \mathcal{P}$ is defined as

$$\operatorname{out}(\alpha, P) = \big\{x \in \operatorname{supp}(\mathcal{P}) : f(x) < K(\alpha)\big\} \tag{6.1}$$

with

$$K(\alpha) = \sup\big\{K > 0 : P\big(\{y : f(y) < K\}\big) \le \alpha\big\}. \tag{6.2}$$

The key element of this definition is the bound $K(\alpha)$. It is the smallest upper bound that yields a probability equal to or just below α if we integrate over the subset of x-values where $f(x) < K(\alpha)$. Figure 6.1 gives examples of outlier regions for some classical distributions. In the case of a continuous distribution like the multivariate normal distribution, the inequality sign in (6.2) can be replaced by the equality sign. However, for discrete situations the outlier region often has a probability of occurrence below the chosen value of α. For example, the 0.1 outlier region of the Poisson distribution with mean value 6 is given by $\{0, 1\} \cup \{11, 12, \ldots\}$, see Fig. 6.1, and has a joint probability of 0.066 as adding any further value from the support to the outlier region would increase this probability above 0.1.

Note that in Definition 6.1 the outlier region is defined as a subset of the union of supports within the considered family of distributions. This is not important for distribution families with support $\mathbb{R}$ throughout, but it is highly relevant for families of distributions where the support depends on unknown parameter(s). Take for example the shifted exponential distribution family with density

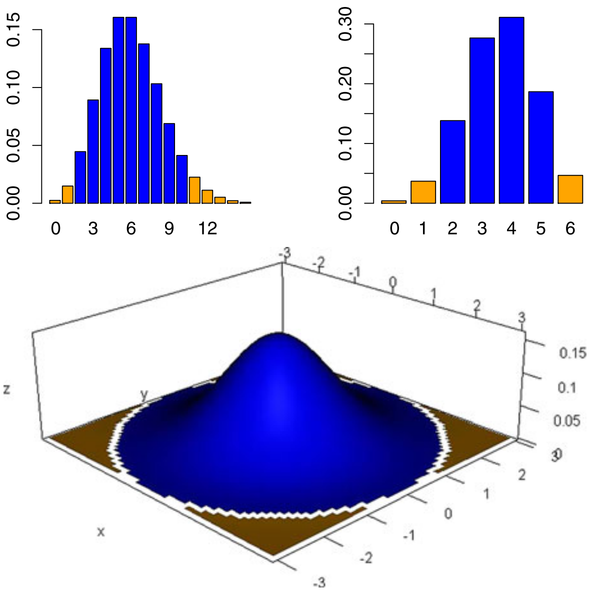


Fig. 6.1 0.1-outlier regions for Poi(6) (*top left*), Bin(6, 0.6) (*top right*) and $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (*bottom*)

$f(x) = \frac{1}{\lambda}\exp(-\frac{x-\theta}{\lambda})1_{[\theta,\infty)}$, $\theta \in \mathbb{R}$, $\lambda > 0$, and support $[\theta, \infty)$. In this case, the α-outlier region is given by the union of the two sets $\{x : x > \theta - \lambda \log \alpha\}$ and $(-\infty, \theta)$. Adding the second set to the α-outlier region has no effect if λ and θ are known, as $(-\infty, \theta)$ has zero probability. However, within outlier detection θ might not be specified correctly or estimated and then it is completely sensible to include values in the outlier set which are possible for other θ values.

The complement of the α-outlier region w.r.t. the support of the distribution is called α-inlier region $\mathrm{inl}(\alpha, P)$. From Definition 6.1, it follows that

$$P\big(X \in \mathrm{inl}(\alpha, P)\big) \geq 1 - \alpha.$$

When dealing with a sample of size N, often α is chosen by taking the sample size into account, i.e., $\alpha_N = 1 - (1 - \alpha)^{1/N}$. Thereby, it is ensured that for $X_1, \ldots, X_N$ i.i.d. according to the model distribution the probability of all observations lying inside the inlier region is at least $1 - \alpha$,

$$P\big(X_i \in \mathrm{inl}(\alpha_N, P), i = 1, \ldots, N\big) \geq 1 - \alpha.$$

Applying Definition 6.1 often yields the tails of a distribution. Gather et al. (2003) discuss various α-outlier regions in the case of uni- or multivariate distributions when $(\mathcal{X}, \mathcal{P}) = (\mathbb{R}^p, \mathcal{B})$, $p \in \mathbb{N}$, where $\mathcal{P} = \{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k\}$ is a family of distributions and each $P_{\boldsymbol{\theta}} \in \mathcal{P}$ has density $f(\cdot, \boldsymbol{\theta})$. Furthermore, $\mathcal{B}$ is the Borel-σ-algebra. W.r.t. a univariate continuous distribution $P_{\boldsymbol{\theta}}$ it is feasible to check whether its α-outlier region coincides with its α- or $\frac{\alpha}{2}$-tail region(s) using typical properties of $P_{\boldsymbol{\theta}}$:

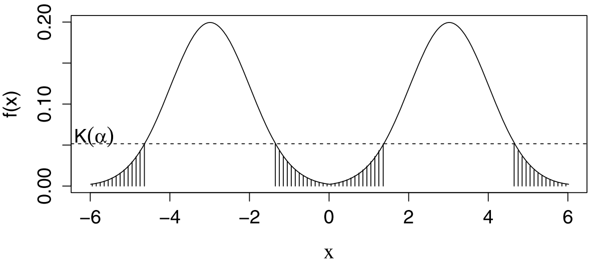


Fig. 6.2 0.1-outlier region (*shaded*) for a bimodal distribution

(i) symmetry:

$$\exists \mu \in \mathbb{R}: \quad f(\mu - x, \boldsymbol{\theta}) = f(\mu + x, \boldsymbol{\theta}) \quad \forall x,$$

(ii) strictly increasing-decreasing density: for some $\mu_1, \mu_2 \in \mathbb{R}$ with $\mu_1 \leq \mu_2$,

$$f(\cdot, \boldsymbol{\theta}) \text{ is strictly increasing on } \operatorname{supp}(P_{\boldsymbol{\theta}}) \cap \big(-\infty, \mu_1], \text{ constant on } [\mu_1, \mu_2] \\ \text{and strictly decreasing on } \operatorname{supp}(P_{\boldsymbol{\theta}}) \cap [\mu_2, \infty\big),$$

(iii) strictly decreasing density:

$$f(\cdot, \boldsymbol{\theta}) \text{ is strictly decreasing on } \operatorname{supp}(P_{\boldsymbol{\theta}}),$$

(iv) strictly increasing density:

$$f(\cdot, \boldsymbol{\theta}) \text{ is strictly increasing on } \operatorname{supp}(P_{\boldsymbol{\theta}}).$$

Thereby we get the following lemma.

Lemma 6.1 (Cf. Gather et al. 2003)

(a) *If $P_{\boldsymbol{\theta}}$ has properties* (i) *and* (ii), *then the corresponding α-outlier region coincides with its lower and upper $\frac{\alpha}{2}$-tail regions.*
(b) *If $P_{\boldsymbol{\theta}}$ has property* (iii), *then the corresponding α-outlier region coincides with its upper α-tail region.*
(c) *If $P_{\boldsymbol{\theta}}$ has property* (iv), *then the corresponding α-outlier region coincides with its lower α-tail region.*

We would like to remark that Lemma 6.1 is very similar to Lemma 2 in Gather et al. (2003), but we added property (ii) because bimodal, symmetric distributions might have an α-outlier region in the center of the distribution, see Fig. 6.2. Furthermore, we extended Lemma 6.1 to distributions with strictly increasing density, like the Beta(3, 1) distribution.

The α-outlier regions for other distributions (continuous ones like χ^2 or Weibull as well as discrete ones like Binomial and Poisson) can be derived by numerical integration (or summation) of the densities. This procedure can also be applied to the multivariate counterparts of those distributions. Let us now consider the general outlier region from (6.1) in case of the normal distribution. Then

$$\text{out}\big(\alpha, \mathcal{N}(\mu, \sigma^2)\big) = \big\{x : |x - \mu| > \sigma z_{1-\alpha/2}\big\},$$

where $z_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$-quantile of the standard normal distribution. The notion can easily be extended to any p-dimensional normal distribution with the help of the relation between the α-outlier regions and certain density contours, explicitly:

$$\text{out}\big(\alpha, \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\big) = \big\{\mathbf{x} \in \mathbb{R}^p : (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) > \chi^2_{p,1-\alpha}\big\}, \tag{6.3}$$

where $\chi^2_{p,1-\alpha}$ denotes the $(1 - \alpha)$-quantile of the χ^2_p distribution. In real data examples the parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ of the given distribution are usually unknown and need to be estimated. Clearly the threshold value $\chi^2_{p,1-\alpha}$ has to be adjusted. Since $(\mathbf{x} - \hat{\boldsymbol{\mu}})' \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}})$ is a Mahanalobis-type distance, for appropriate estimators $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ one can use asymptotics to derive $\chi^2_{p,1-\alpha_N}$ as the new threshold value. Consider Example 6.1, p. 91 as an illustration. The contribution by Becker, Liebscher and Kirschstein (in Chap. 7) deepens the discussion of outliers in multivariate cases. Furthermore, there exist more complex data situations than simple uni- or multivariate raw data tables, some of which we discuss in more detail below.

6.3 Detection of α-Outliers

Generally we can distinguish between identification rules that identify α-outliers in one-step and stepwise rules that successively judge the outlyingness of observations. In a stepwise outward procedure, one first defines a subset of observations taken to be free of outliers. Then the "least conspicuous" or in some sense "least outlying" observations are successively tested for being outliers. If not judged as too outlying, they are added to the current subset. Note that here we may judge "least outlyingness" with respect to α-outlier regions derived from the current subset. Inward procedures work similar starting from the full data set from which observations identified as outliers are successively removed.

Although stepwise procedures have been considered within the framework of identifying α-outliers (e.g., Davies and Gather 1993; Schultze and Pawlitschko 2000), the main interest lies within one-step rules. Roughly spoken, based on some empirical outlier regions all observations lying within this region are identified as outliers in a single step. Such rules have various advantages over step-wise rules, besides the most important one of frequently showing a better performance with respect to the task of outlier detection (see, e.g., Davies and Gather 1993; Kuhnt 2004).They are relatively easy to apply and interpret. Parameter estimates only need to be calculated once from the full set and not successively from subsets for which they might not exist or not be unique anymore.

Let us consider the following outlier identification problem. Here $(\mathbf{X}_1, \ldots, \mathbf{X}_N)$ denotes a sample with corresponding observations $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$. The data generation mechanism is assumed to follow a distribution $P_{\boldsymbol{\theta}}$ from a class $\mathcal{P} = \{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ for the regular observations. Now we want to identify observations within the sample which lie in the α-outlier region of $P_{\boldsymbol{\theta}}$.

As $\boldsymbol{\theta}$ is unknown, we naturally start by estimating $\boldsymbol{\theta}$ on the sample. With the estimate $\hat{\boldsymbol{\theta}}$ we can build an empirical outlier region $\mathrm{out}(c(\alpha, N), P_{\hat{\boldsymbol{\theta}}}) = \mathrm{out}(c(\alpha, N), \hat{P})$, where c denotes a constant usually chosen depending on α and the sample size N. We now may identify observations in this thereby estimated outlier region as outliers, leading us towards a one-step outlier identification rule, next defined in more general terms.

Definition 6.2 Let an empirical outlier region $\mathrm{out}(c(\alpha, N), \hat{P})$ be given. A mapping OIF (from the support $\mathrm{supp}(\mathcal{P})$ of the considered class of distributions to the set $\{0, 1\}$), $\mathrm{OIF} : \mathrm{supp}(\mathcal{P}) \to \{0, 1\}$, given by

$$\mathrm{OIF}\big(\mathbf{x} \mid (\mathbf{X}_1, \ldots, \mathbf{X}_N), \alpha\big) = \mathbf{1}_{\mathrm{out}(c(\alpha,N),\hat{P})}(\mathbf{x}), \quad \mathbf{x} \in \mathrm{supp}(\mathcal{P}),$$

with the interpretation

$$\mathrm{OIF}\big(\mathbf{x} \mid (\mathbf{X}_1, \ldots, \mathbf{X}_N), \alpha\big) = \begin{cases} 1, & \mathbf{x} \text{ is identified as outlier,} \\ 0, & \mathbf{x} \text{ is not identified as outlier,} \end{cases}$$

is called a *one-step outlier identification rule*.

As Gather et al. (2003) state: "(…) Outliers in the data may seriously affect standard estimators of unknown distribution parameters", it is recommended to use robust estimators within the one-step outlier identification. Depending on the data situation, typical robust estimators chosen within outlier detection are the median, L_1, M-estimators like Huber or Hampel or any other estimator with a higher breakdown point (see the contribution of Rousseeuw and Hubert in Chap. 4, or Schultze and Pawlitschko 2002). One-step outlier identification rules in the sense of Definition 6.2 have for example been considered for the univariate and multivariate normal distribution (see, e.g., Becker and Gather 1999) and the exponential distribution (see, e.g., Schultze and Pawlitschko 2002). In the case of more structured data situations like one-way random effect models (Wellmann and Gather 2003), regression models (Boscher 1992), logistic regression (Christmann 1992), time series models (Gather et al. 2002) as well as contingency tables (Kuhnt 2004) adjustments to this basic one-step procedure are sometimes necessary, some of which are discussed later.

The constant c is often fixed by applying some general normalizing condition such as

$$P(\text{“no outliers identified in the sample”}) = 1 - \alpha$$

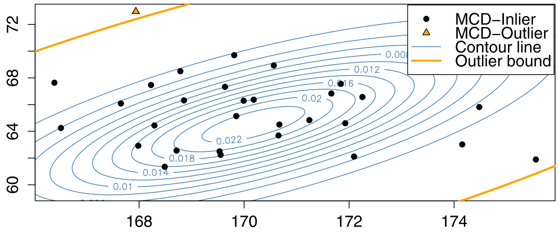


Fig. 6.3 Contour plot and 0.01-outlier region for a bivariate normal distribution, estimated by FAST MCD

or

$$P(\text{“the empirical outlier region lies within the true outlier region”}) = 1-\alpha.$$

In rare cases, the normalizing constant c can be calculated exactly. However, more often it needs to be simulated and may even depend on the unknown parameter as well as the sample size and the specific identification rule. Therefore sometimes just $c=\alpha_N$ or even $c=\alpha$ is used.

The performance of outlier identification rules is typically measured by criteria like the largest non-identifiable outlier, as well as masking and swamping breakdown points (Kuhnt 2010). Roughly spoken, breakdown points are then given by the minimal fraction of nonregular observations needed to cause the effect. Tietjen and Moore (1972) state that the “masking effect is the inability (...) to identify even a single outlier in the presence of several suspected values.” The identification procedure is manipulated by the outliers, which mask themselves. On the other hand outliers can cause that true inliers are identified as outliers, this effect is called swamping.

Example 6.1 (Multivariate Normal Distribution) Consider the size and weight of 30 female students, displayed in Fig. 6.3, for which we assume that the regular observations follow a multivariate normal distribution. After computing the FAST MCD estimator (see Rousseeuw 1984; Rousseeuw and van Driessen 1999, and the contribution by Rousseeuw and Hubert in Chap. 4) of mean $\hat{\boldsymbol{\mu}}_{\text{MCD}} = (170.178, 64.525)'$ and covariance

$$\hat{\boldsymbol{\Sigma}}_{\text{MCD}} = \begin{pmatrix} 6.955 & 6.621 \\ & 12.849 \end{pmatrix},$$

we compute the 0.01-outlier region and check for outliers.

The outlier bound of the FAST MCD estimator is given in Fig. 6.3. The FAST MCD estimator identifies one value as outlying (orange triangle), a student with high weight (72.9 kg) compared to her height (167.9 cm). Calculation of the respective ML-estimates give $\hat{\boldsymbol{\mu}}_{\text{ML}} = (170.117, 65.481)'$, $\hat{\boldsymbol{\Sigma}}_{\text{ML}} = \left(\begin{smallmatrix} 4.909 & -1.363 \\ & 7.126 \end{smallmatrix}\right)$ and no observation lies in the 0.01-outlier region of the normal distribution with these parameters. We observe that we get higher variance/covariance values which prevent the identification of outliers, such that we might have a masking effect. Plus, the ML-estimates yield a questionable negative correlation between these variables.

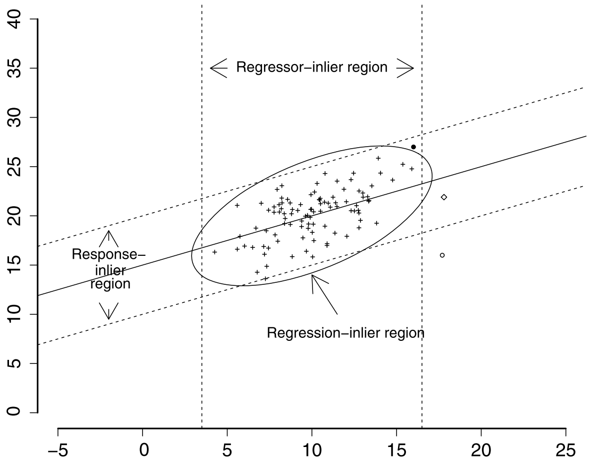


Fig. 6.4 Different types of outliers in the regression context

6.4 Outliers in Regression

In the simple regression context there are three different ways to think of α-outliers: regressor-, response- and regression-outliers (Fig. 6.4). For example, response outliers are presented in Boscher (1992) and regression outliers are treated generally in Rousseeuw and Leroy (1987).

First of all, we consider the very intuitive response-outlier (see also Boscher 1992). Each response value deviating from the regression line by more than some specific constant will turn out to be a response-α-outlier. As highlighted in the previous sections, we need assumptions w.r.t. the distribution of the response for the calculation of α-outlier regions. Considering a simple linear regression model

$$Y = \beta_0 + \mathbf{X}'\boldsymbol{\beta}_1 + U,$$

it is commonly assumed that the conditional distribution of the response given the regressor vector is normal:

$$P_{Y|\mathbf{X}} = \mathcal{N}\big(\beta_0 + \mathbf{X}'\boldsymbol{\beta}_1, \sigma^2\big)$$

for a scale parameter $\sigma^2 > 0$. The residuals $U_i, i = 1, \dots, n$, are assumed to be normally distributed given the regressor vector $\mathbf{X}$ with $E(U) = 0$ and $\text{Var}(U) = \sigma^2$. Hence, the corresponding response-α-outlier region can be defined as

$$\text{out}(\alpha, P_{Y|\mathbf{X}}) = \big\{y \in \mathbb{R} : u = \big|y - \big(\beta_0 + \mathbf{X}'\boldsymbol{\beta}_1\big)\big| > \sigma z_{1-\alpha/2}\big\}. \tag{6.4}$$

This method is a neat way for regression setups where the regressors themselves are non-random, especially if a statistical design of experiment was used. If the regressors are also stochastic, a common assumption is a p-variate normal distribution

$$P_{\mathbf{X}} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{6.5}$$

with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and $(p \times p)$-covariance matrix $\boldsymbol{\Sigma}$. It is of course also possible to calculate α-outlier regions for the regressors. The outlier region depends on the distance between the observation and the mean and on the covariance matrix:

$$\operatorname{out}(\alpha, P_{\mathbf{X}}) = \left\{\mathbf{x} \in \mathbb{R}^p : (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) > \chi^2_{p,1-\alpha}\right\}, \tag{6.6}$$

which resembles equation (6.3) because of (6.5). The notion of observations with a high impact on the regression (Rousseeuw and van Zoomeren 1990), so-called leverage points, can be formalized by (6.6).

The third way to define an α-outlier in the regression context is by taking the joint distribution of Y and $\mathbf{X}$ into account, resulting in an elliptically shaped contour which resembles the intersection of the two previously mentioned α-inlier regions. This yields the regression-α-outlier region, defined as

$$\begin{aligned}\operatorname{out}(\alpha, P_{(Y,\mathbf{X})}) = \Big\{\big(y, \mathbf{x}'\big)' \in \mathbb{R}^{p+1} : \big(y - \big(\beta_0 + \mathbf{x}'\boldsymbol{\beta}_1\big)\big)^2/\sigma^2 \\ + (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) > \chi^2_{p+1,1-\alpha}\Big\}.\end{aligned} \tag{6.7}$$

We use a similar set-up as in Gather et al. (2003) to exemplify the different types of outliers in Fig. 6.4, with $\mu = 10$, $\beta_0 = 15$, $\beta_1 = 1/2$, $\sigma = \sqrt{6}$, $\Sigma = \sqrt{6}$ and $N = 100$ observations. Now the observed point denoted by "•" lies inside the regressor-inlier region as well as in the response-inlier region but outside the regression-inlier region and therefore is only a regression outlier whereas "○" is an outlier with respect to all regions. The third point outside the regression-inlier region ◇ is a regressor-outlier but no response-outlier as it lies very close to the regression line.

Now, a one-step outlier identification method for the regression case can easily be derived based on robust estimators of $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, σ and $\boldsymbol{\Sigma}$, respectively, and the resulting outlier regions of the thereby estimated distributions.

Example 6.2 (Thermal Spraying) Consider a thermal spraying process in the coating industry, wherein particles are sprayed onto some material. We study the effect of the temperature and velocity of these particles on the coating property porosity. The results of 30 runs (Table 6.1) are analyzed using a main effect quantile regression with $\tau = 0.5$: $\hat{\boldsymbol{\beta}} = (52.997, -0.014, -0.035)'$.

Robust estimation of the parameters of $P_{\mathbf{X}}$ with the MCD (see the contribution by Rousseeuw and Hubert in Chap. 4) yields $\hat{\boldsymbol{\mu}}_{\text{MCD}} = (1517.650, 715.169)'$ and

$$\hat{\boldsymbol{\Sigma}}_{\text{MCD}} = \begin{pmatrix} 8292.525 & -1949.700 \\ & 1737.459 \end{pmatrix}.$$

We are interested in possible response- or regression-outliers. Application of (6.4) using the above estimates and $c(\alpha, N) = 0.05$ yields no response-outliers. We next use (6.7) with three degrees of freedom and $\alpha = 0.05$ to estimate a regression-outlier region. Note that we abstain from applying any standardization of the identifier w.r.t. the null model as well as an adjustment of α to the sample size. In practice, this simple approach has turned out to be sufficient for most purposes. The one-step procedure yields four α-outliers, namely run 8, 9, 13 and 30. Run 8 is represented by the highest temperature and one of the highest velocities. From the two negative regres-

Table 6.1 Data set: thermal spraying example

Runs 1–10			Runs 11–20			Runs 21–30		
Porosity	Temperature	Velocity	Porosity	Temperature	Velocity	Porosity	Temperature	Velocity
5.86	1525.50	685.20	4.82	1563.00	715.80	10.54	1448.70	710.20
3.31	1621.50	749.60	4.91	1626.50	763.60	9.60	1485.40	727.90
6.64	1562.40	658.90	**3.29**	**1598.60**	**791.50**	6.54	1493.70	701.30
8.29	1605.20	645.70	6.18	1619.40	743.00	7.48	1416.50	754.90
4.18	1606.70	695.00	7.97	1498.10	673.50	8.29	1480.20	742.10
4.74	1562.20	726.40	9.46	1532.50	644.10	10.06	1455.50	753.90
5.99	1618.00	712.00	8.80	1565.20	678.30	7.40	1449.40	728.80
5.53	**1669.70**	**765.60**	3.96	1517.40	736.40	4.01	1511.70	792.10
3.07	**1629.20**	**786.30**	10.69	1550.20	715.70	7.13	1492.10	720.60
8.06	1548.90	721.00	7.24	1538.30	684.10	**11.01**	**1404.20**	**647.60**

sion coefficients given above one would expect to have one of the smallest values for porosity, but it turns out to be quite average. The other three outliers are characterized by very small or high values of porosity, especially if the velocity values are taken into account. Taking this information into account helps us to understand the thermal spraying process better—a possible conclusion is that the quantile regression does not perform very well with extreme regressor values. In fact, Rehage et al. (2012) show that gamma generalized linear models outperform regression in this kind of process.

6.5 Outliers in Contingency Tables

Not all data used and analyzed by scientists are continuous. Especially in the field of social sciences one has to deal with categorical data. Often these kind of data are presented as contingency tables. Let $\mathbf{X}_\Delta = (X_1, \dots, X_p)'$ be a p-dimensional random vector with components $X_\delta, \delta \in \Delta = \{1, \dots, p\}$. Each X_δ is a categorical random variable with I_δ possible outcomes. A sample of N observations $(x_1^1, \dots, x_p^1)', \dots, (x_1^N, \dots, x_p^N)'$ is the (data) basis for a contingency table. The support of the random vector $\mathbf{X}_\Delta$ is given by the set $\mathcal{I} = \times_{\delta=1}^{p} \{1, \dots, I_\delta\}$, $|\mathcal{I}| = I$. The cells of a contingency table are determined by $\mathcal{I}$ and contain the number of times n_i each combination occurs in the data set, $i = 1, \dots, I$, which are understood as realizations of random variables N_i. To apply the concept of α-outliers to contingency tables, we need a model for the cell counts of the table. There are two widely used assumptions depending on whether the sample size N is fixed a priori or not. In the first case, we can assume a multinomial distribution for the vector $(N_i)_{i \in \mathcal{I}}$, in the second case the N_i are assumed to follow independent Poisson distributions with parameters m_i, $i \in \mathcal{I}$. In both cases, loglinear models are used to model the independence structure between the original variables $X_1, \dots, X_p$ (Bishop et al. 1975).

An important issue with respect to outlier detection is whether the whole contingency table is to be checked as outlying or only each cell of the contingency table. As the presence of more than one contingency table is rather seldom, outlier detection in contingency tables mostly concentrates on single outlying cells (Simonoff 1988; Fuchs and Kenett 1980; Upton and Guillen 1995; Kuhnt 2004).

6.5.1 Outliers in Multinomial Models

In the case of a multinomial model, the concept of α-outliers would apply to the complete table as the vector of all cell counts follows a multinomial distribution. This might be of interest if there is a couple of contingency tables, especially if they are collected by different authorities. The researcher can check whether one or more of these contingency tables can be called “outlying”. This might for example be the case if the same questionnaire has been used to interview the same number of people in different cities, resulting in a contingency table for each city.

If we want to have an outlier definition based on the α-outlier idea for individual cell counts, we can refer to the marginal binomial distribution.

Using the one-step outlier identifier approach a cell count n_i is identified as outlier if it lies in $\operatorname{out}(c(\alpha, N), \operatorname{Bin}(\hat{p}_i, N))$, where $\hat{p}_i$ denotes an estimate of the cell probability. Here again the use of a robust estimator is recommendable, e.g., the so-called Pearson least trimmed chi-squared residual estimator (LTCS, see Shane and Simonoff (2001)). Note, however, that the product of the marginal distributions is not consistent with the assumed distribution for the complete vector of cell counts.

6.5.2 Outliers in Poisson Models

In loglinear Poisson models the vector of parameters of the individually Poisson distributed random variables N_i is given by a parameter space $\mathcal{M} \subseteq \mathbb{R}^I$, i.e., $m_i, m_{\mathcal{I}} = (m_i)_{i \in \mathcal{I}} \in \mathcal{M}$ (Bishop et al. 1975). The unknown parameter vector is usually estimated by the maximum-likelihood method, robust alternatives like the L_1-estimator are so far rarely considered. One-step outlier identification procedures in the sense of Definition 6.2 are then given as

$$\operatorname{OIF}\bigl(n_i; n_{\mathcal{I}}, c(\alpha, N)\bigr) = \mathbf{1}_{\operatorname{out}(c(\alpha,N),\hat{m}_i(n_{\mathcal{I}}))}(n_i), \quad i \in \mathcal{I},\ n_{\mathcal{I}} \in \mathbb{N}^I. \tag{6.8}$$

Kuhnt (2004) compares one-step procedures and outward procedures using maximum-likelihood estimates as well as L_1-estimates by a simulation study. The one-step procedure based on L_1-estimates outperforms the other three procedures in nearly all of the treated outlier situations.

Example 6.3 (Students’ Subjects) We are interested in data from 88 students collected in their first statistics lesson at TU Dortmund University. They were asked

Table 6.2 Contingency table and inlier regions of the students' subjects example

	St	DA	ME	CS	Ma	Ph	Ps	Σ
within 20 km								
n_i	12	6	2	1	5	4	4	34
$\hat{m}_i$	16.96	6.00	2.72	2.72	4.27	4.65	4.65	
inl(0.1, Poi($\hat{m}_i$))	$\{10,\dots,23\}$	$\{2,\dots,10\}$	$\{0,\dots,5\}$	$\{0,\dots,5\}$	$\{1,\dots,7\}$	$\{1,\dots,8\}$	$\{1,\dots,8\}$	
farther away								
n_i	**29**	7	2	2	2	6	6	54
$\hat{m}_i$	19.79	7.00	3.17	3.17	4.98	5.43	5.43	
inl(0.1, Poi($\hat{m}_i$))	$\{13,\dots,27\}$	$\{3,\dots,11\}$	$\{1,\dots,6\}$	$\{1,\dots,6\}$	$\{2,\dots,9\}$	$\{2,\dots,9\}$	$\{2,\dots,9\}$	

whether their hometown is within a radius of 20 km of the university or not (X_1: Distance) and which subject they study (X_2: Subject). Apparently, we can expect many students of statistics (coded St) in this lesson. But also students of other subjects attend the lesson, like data analysis (DA), mathematical economics (ME), computer sciences (CS), mathematics (Ma), physics (Ph) and psychology (Ps), see Table 6.2.

We consider the loglinear Poisson model based on the independence assumption of the two original variables X_1 and X_2. Computing the α-outlier regions based on L_1-estimates yields one outlier: the students whose hometown is far away and who study statistics now. Here, for simplicity we choose $c(\alpha, N) = \alpha = 0.1$. The estimated α-inlier region is given by $\{13,\dots,27\}$. This underlines the point that a subject which can be studied only at a small number of universities (like statistics in Germany) attracts potential students from a bigger radius than other subjects.

6.6 Outliers in Graphical Models

Graphical models (Lauritzen 1996) are an interesting way to visualize the dependency structure of a data set with a large number of variables, especially when both continuous and discrete variables are considered. Even moderate outliers may contaminate the estimated dependency structure such that a reasonable interpretation becomes impossible, see Kuhnt and Becker (2003). As pointed out in Vogel and Fried (2010), robust estimation of multivariate scatter is a very useful tool, especially if one aims at detecting outliers. For the application to the concept of α-outliers, we need a distributional assumption of graphical models, where we concentrate on observations from random vectors with continuous as well as categorical components.

First of all, we fix some notations: Let $\mathbf{X} = (\mathbf{X}'_\Delta, \mathbf{X}'_\Gamma)'$ be a $(p+q)$-dimensional random vector with p discrete variables $\mathbf{X}_\delta, \delta \in \Delta$ and q continuous variables $\mathbf{X}_\gamma, \gamma \in \Gamma$. The discrete random vector will take values from $\mathcal{I} = \bigotimes_{\delta\in\Delta} I_\delta$ with I_δ as the range of values $\mathbf{X}_\delta$ can take. Lauritzen and Wermuth (1989) define a distribution, where the continuous variables given the discrete variables follow a Gaussian distribution, the so-called conditional Gaussian distribution (CG-distribution).

Definition 6.3 (Lauritzen and Wermuth 1989) A random vector $\mathbf{X} = (\mathbf{X}'_\Delta, \mathbf{X}'_\Gamma)'$ follows a conditional Gaussian distribution, if and only if its density can be written as

$$f_{\mathbf{X}_\Delta,\mathbf{X}_\Gamma}(\mathbf{i},\mathbf{y}) = f_{\mathbf{X}_\Delta}(\mathbf{i}) f_{\mathbf{X}_\Gamma|\mathbf{X}_\Delta}\big(\mathbf{y} \mid \mathbf{i}\big),$$

yielding

$$f_{\mathbf{X}_\Delta,\mathbf{X}_\Gamma}(\mathbf{i},\mathbf{y}) = p(\mathbf{i})(2\pi)^{-q/2}\det\big(\mathbf{\Sigma}(\mathbf{i})^{-1/2}\big)\exp\big\{-\big(\mathbf{y}-\boldsymbol{\mu}(\mathbf{i})\big)'\mathbf{\Sigma}(\mathbf{i})^{-1}\big(\mathbf{y}-\boldsymbol{\mu}(\mathbf{i})\big)/2\big\},$$

where $p(\mathbf{i})$ is the probability of the occurrence of $\mathbf{i}$ and $\boldsymbol{\mu}(\mathbf{i})$, $\mathbf{\Sigma}(\mathbf{i})$ are the conditional mean and covariance of $\mathbf{X}_\Gamma$ given $\mathbf{X}_\Delta = \mathbf{i}$.

Given Definition 6.3 we can apply (6.1) to CG-distributions:

Definition 6.4 (Kuhnt 2006) The α-outlier region with respect to a given conditional Gaussian distribution with density $f(\mathbf{X}_\Delta = \mathbf{i}, \mathbf{X}_\Gamma = \mathbf{y}) = p(\mathbf{i}) f(\mathbf{y} \mid \mathbf{i})$ is defined by

$$\operatorname{out}(\alpha, P) = \big\{(\mathbf{i},\mathbf{y}) \in \mathcal{I} \times \mathbb{R}^{|\Gamma|} : p(\mathbf{i}) f\big(\mathbf{y} \mid \mathbf{i}\big) < K(\alpha)\big\}, \tag{6.9}$$

where

$$K(\alpha) = \sup\big\{K > 0 : P\big(\big\{\big(\mathbf{i}^*,\mathbf{y}^*\big) : p\big(\mathbf{i}^*\big) f\big(\mathbf{y}^* \mid \mathbf{i}^*\big) < K\big\}\big) \le \alpha\big\}.$$

To exemplify this definition, we derive the α-outlier region of a CG-distribution with only one continuous random variable, i.e., $q = 1$ from (6.9) in detail:

$$\begin{aligned}
&p(\mathbf{i})\frac{1}{\sqrt{2\pi}\sigma(\mathbf{i})}\exp\left(-\frac{(y-\mu(\mathbf{i}))^2}{2\sigma(\mathbf{i})^2}\right) < K(\alpha)\\
&\Leftrightarrow y > \sqrt{-2\sigma(\mathbf{i})^2\ln\left(\frac{K(\alpha)\sqrt{2\pi}\sigma(\mathbf{i})}{p(\mathbf{i})}\right)} + \mu(\mathbf{i})\\
&\quad \vee\, y < -\sqrt{-2\sigma(\mathbf{i})^2\ln\left(\frac{K(\alpha)\sqrt{2\pi}\sigma(\mathbf{i})}{p(\mathbf{i})}\right)} + \mu(\mathbf{i}).
\end{aligned}$$

The conditional probability of this event given $\mathbf{i}$ is:

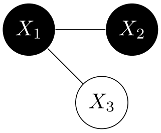
$$\begin{aligned}
&P_{\mathbf{X}_\Gamma|\mathbf{X}_\Delta=\mathbf{i}}\left(\mathbf{X}_\Gamma > \sqrt{-2\sigma(\mathbf{i})^2\ln\left(\frac{K(\alpha)\sqrt{2\pi}\sigma(\mathbf{i})}{p(\mathbf{i})}\right)} + \mu(\mathbf{i}) \vee \mathbf{X}_\Gamma\right.\\
&\left. < -\sqrt{-2\sigma(\mathbf{i})^2\ln\left(\frac{K(\alpha)\sqrt{2\pi}\sigma(\mathbf{i})}{p(\mathbf{i})}\right)} + \mu(\mathbf{i})\right)\\
&= 2\Phi\left(-\sqrt{-2\ln\left(\frac{K(\alpha)\sqrt{2\pi}\sigma(\mathbf{i})}{p(\mathbf{i})}\right)}\right),
\end{aligned}$$

where Φ is the standard normal distribution function. The overall probability for the α-outlier region of the CG-distribution is therefore

$$\sum_{\mathbf{i}\in\mathcal{I}} p(\mathbf{i}) 2\Phi\left(-\sqrt{-2\ln\left(\frac{K(\alpha)\sqrt{2\pi}\sigma(\mathbf{i})}{p(\mathbf{i})}\right)}\right) \leq \alpha. \tag{6.10}$$

Using this result, it is now possible to determine $K(\alpha)$ numerically given α and the needed parameter values.

Example 6.4 (Graphical Model)



A simple case to illustrate α-outlier regions w.r.t. CG-distributions is the appearance of two discrete variables X_1, X_2 with two possible outcomes each ($\mathcal{I} = \{(1,1), (1,2), (2,1), (2,2)\}$) and one continuous variable X_3. Let the probability vector of $(X_1, X_2)'$ be given by

$$\big(p(11), p(12), p(21), p(22)\big)' = (0.2, 0.01, 0.3, 0.49)'$$

and the parameters of the conditional densities of $X_3 \mid (X_1, X_2)' = \mathbf{i}$ by

$$\big(\mu(11), \mu(12), \mu(21), \mu(22)\big)' = (0, 0, 1, 1)'$$

and

$$\big(\sigma(11), \sigma(12), \sigma(21), \sigma(22)\big)' = (1, 1, 4, 4)'.$$

Equation (6.10) yields $K(0.1) = 0.0103099$ for this CG-distribution P, therefore the outlier regions can be derived with (6.9). The calculation of the α-outlier region of $X_3 \mid ((X_1, X_2)' = (1, 2)')$ is not needed because the probability of this event ($p(12) = 0.01$) is smaller than $K(0.1)$. Therefore the whole support of X_3 coincides in this case with the α-outlier region. The 0.1-outlier region is given by

$$\begin{aligned}\mathrm{out}(0.1, P) = \Big\{&\big\{(1,1,y) : y < -2.023 \vee y > 2.023\big\}\\ &\cup\big\{(1,2,y) : y \in \mathbb{R}\big\}\\ &\cup\big\{(2,1,y) : y < -4.839 \vee y > 6.839\big\}\\ &\cup\big\{(2,2,y) : y < -5.056 \vee y > 8.056\big\}\Big\},\end{aligned}$$

see Fig. 6.5.

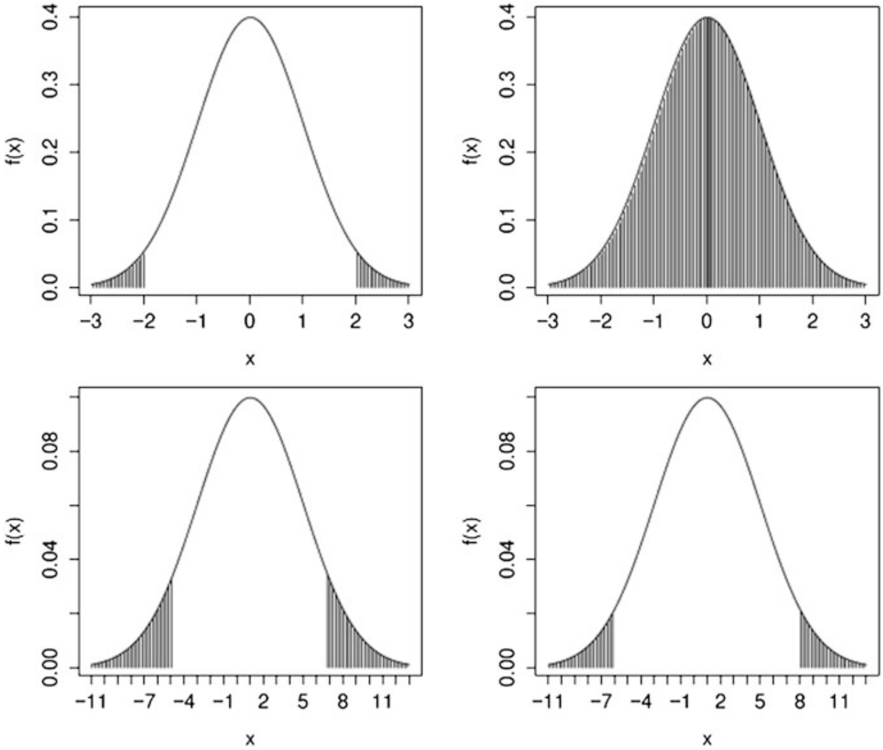


Fig. 6.5 0.1-outlier regions (*shaded*) of Example 6.1. *Top left*: $p(11) = 0.2, \mu(11) = 0$, $\sigma^2(11) = 1$; *top right*: $p(12) = 0.01, \mu(12) = 0, \sigma^2(12) = 1$; *bottom left*: $p(21) = 0.3$, $\mu(21) = 1, \sigma^2(21) = 16$; *bottom right*: $p(22) = 0.49, \mu(22) = 1, \sigma^2(22) = 16$

6.7 Conclusions

The concept of α-outliers is an impartial way to identify observations which deviate from the bulk of the data. The smaller the chosen α, the more conservative the outlier detection becomes. This concept is applicable in any model-based context. Usually the parameter vector of the assumed distribution is unknown and therefore has to be estimated in advance. Here it is important to use robust estimators as otherwise the estimates might be contaminated by potential outliers and might cause the effects of masking and swamping.

We presented a number of structured data situations where α-outliers can be applied. Of course, further situations exist where the computation of α-outliers is feasible. The identification of outliers in online monitoring data is treated in Gather et al. (2002). Wellmann and Gather (2003) discuss an application of α-outliers in a one-way random effects model. The identification of α-outliers in logistic regression is explored extensively by Christmann (1992).

Acknowledgements The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, project A1 and SFB 823, project B1) is gratefully acknowledged.

References

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data* (3rd ed.). Chichester: Wiley & Sons.

Becker, C., & Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association*, *94*, 947–955.

Becker, C., & Gather, U. (2001). The largest nonidentifiable outlier: A comparison of multivariate simultaneous outlier identification rules. *Computational Statistics & Data Analysis*, *36*, 119–127.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis*. Cambridge: MIT Press.

Boscher, H. (1992). *Behandlung von Ausreißern in linearen Regressionsmodellen*. Dissertation, Universität Dortmund.

Christmann, A. (1992). *Ausreißeridentifikation und robuste Schätzer im logistischen Regressionsmodell*. Dissertation, Universität Dortmund.

Davies, P. L., & Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association*, *88*, 782–792.

Fuchs, C., & Kenett, R. (1980). A test for detecting outlying cells in the multinomial distribution and two-way contingency tables. *Journal of the American Statistical Association*, *75*, 395–398.

Gather, U., Bauer, M., & Fried, R. (2002). The identification of multiple outliers in online monitoring data. *Estadística*, *54*, 289–338.

Gather, U., Kuhnt, S., & Pawlitschko, J. (2003). Concepts of outlyingness for various data structures. In J. C. Misra (Ed.), *Industrial Mathematics and Statistics* (pp. 545–585). New Dehli: Narosa Publishing House.

Kuhnt, S. (2004). Outlier identification procedures for contingency tables using maximum likelihood and L_1 estimates. *Scandinavian Journal of Statistics*, *31*, 431–442.

Kuhnt, S. (2006). Robust graphical modelling for mixed variables.

Kuhnt, S. (2010). Breakdown concepts for contingency tables. *Metrika*, *71*, 281–294.

Kuhnt, S., & Becker, C. (2003). Sensitivity of graphical modeling against contamination. In M. Schader, W. Gaul, & M. Vichi (Eds.), *Between Data Science and Applied Data Analysis* (pp. 279–287). Berlin: Springer.

Kuhnt, S., & Pawlitschko, J. (2005). Outlier identification rules for generalized linear models. In D. Baier & K.-D. Wernecke (Eds.), *Innovations in Classification, Data Science, and Information Systems* (pp. 165–172). Berlin: Springer.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.

Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, *17*, 31–57.

Rehage, A., Rudak, N., Hussong, B., Kuhnt, S., & Tillmann, W. (2012). *Prediction of in-flight particle properties in thermal spraying with additive day-effects*. Discussion Paper 06/12, SFB 823, TU Dortmund University.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*, 871–880.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Rousseeuw, P. J., & van Zoomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, *85*, 633–639.

Rousseeuw, P. J., & van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*, 212–223.

Schultze, V., & Pawlitschko, J. (2000). *Identification of outliers in exponential samples with stepwise procedures*. Technical Report 56/00, SFB 475, Universität Dortmund.

Schultze, V., & Pawlitschko, J. (2002). The identification of outliers in exponential samples. *Statistica Neerlandica*, *56*, 41–57.

Shane, K. V., & Simonoff, J. S. (2001). A robust approach to categorical data analysis. *Journal of Computational and Graphical Statistics*, *10*, 135–157.

Simonoff, J. S. (1988). Detecting outlying cells in two-way contingency tables via backwards-stepping. *Technometrics*, *30*, 339–345.

Tietjen, G. L., & Moore, R. H. (1972). Testing for a single outlier in simple linear regression. *Technometrics*, *15*, 583–597.

Upton, G. J. G., & Guillen, M. (1995). Perfect cells, direct models and contingency table outliers. *Communications in Statistics. Theory and Methods*, *24*, 1843–1862.

Vogel, D., & Fried, R. (2010). On robust Gaussian graphical modelling. In L. Devroye, B. Karasözen, M. Kohler, & R. Korn (Eds.), *Recent Developments in Applied Probability and Statistics* (pp. 155–182). Berlin: Springer.

Wellmann, J., & Gather, U. (2003). Identification of outliers in a one-way random effects model. *Statistical Papers*, *44*, 335–348.

Chapter 7
Multivariate Outlier Identification Based on Robust Estimators of Location and Scatter

Claudia Becker, Steffen Liebscher, and Thomas Kirschstein

7.1 Introduction

When dealing with real-life data, analysts and researchers are well aware that often there are anomalies, like single observations not fitting the "main body" of the data, clusters of deviating observations forming some different pattern than all other data, etc. It can be assumed that on average 1 %–10 % of the observations in a data set may be extremely deviating (Hampel et al. 1986, p. 28). Analyzing such data sets just with standard statistical methods can yield biased results. Hence, either their identification followed by elimination or the use of robust methods is recommended, see, e.g., the contributions in this book by Borowski, Fried and Imhoff (Chap. 12), Filzmoser and Hron (Chap. 8), Galeano and Peña (Chap. 15), Huskova (Chap. 11), Kharin (Chap. 14), Oja (Chap. 1), Rousseeuw and Hubert (Chap. 4) and Spangl and Dutter (Chap. 13).

Moreover, the unusual observations themselves may be of a certain interest. Sometimes they contain information on special events during the period of data collection, or hints on valuable specialties of a certain topic. A simple but important example is the case of the Chernobyl catastrophe, where extremely high measurements of radioactivity (unusual observations within the usual plant radioactivity data of a nuclear power plant in Sweden) were indicating that something had happened (Mara 2011, p. 50). Hence, detecting such unusual observations can be seen as one step within the data analysis procedure but also as an important task in itself. To be able to fulfill this task, it is necessary to define "unusual" in a first step. There

C. Becker (✉) · S. Liebscher · T. Kirschstein
Martin-Luther-University Halle-Wittenberg, 06099 Halle, Germany
e-mail: claudia.becker@wiwi.uni-halle.de

S. Liebscher
e-mail: steffen.liebscher@wiwi.uni-halle.de

T. Kirschstein
e-mail: thomas.kirschstein@wiwi.uni-halle.de

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_7, © Springer-Verlag Berlin Heidelberg 2013

exist formal definitions with respect to some target distribution, where it is assumed that the majority of the data stems from this target distribution (see the contribution by Kuhnt and Rehage, Chap. 6, for the approach of α outliers). Regions far away from the main part of the distribution (in the sense of having low density and, hence, low probability of being reached by observations generated from the target) are associated as outlier regions.

Once "outliers" are well defined, there is the need to develop methods for identifying them. This is not an easy task, since outliers themselves tend to disturb statistical methods. Well-known effects in this context are masking (outliers can not be found because they are hidden for the detection method by other outliers or by themselves, see, e.g., Pearson and Chandra Sekar 1936; Murphy 1951; Barnett and Lewis 1994; Becker and Gather 1999; Dang and Serfling 2010) and swamping (because of outliers in the data observations that are no outliers at all are falsely identified as deviating, see, e.g., Fieller 1976; Davies and Gather 1993). Often, using robust statistical procedures within outlier identification rules can reduce these problems, although usually assertions on bounded masking and/or swamping effects are based on asymptotic behavior of the procedures (Davies and Gather 1993; Becker and Gather 1999, amongst others, also see the contributions by Oja, Chap. 1, and by Rousseeuw and Hubert, Chap. 4, for robust measures of location and scatter to use within such procedures). In finite samples, even for "good" methods with respect to these effects still the size of the largest non-identifiable outlier can be quite large (Becker and Gather 2001).

However, not in all situations even exists a consensus of what constitutes an outlier. In particular, if the knowledge about the target distribution is rather vague, it might not be immediately clear how to define "outlying". In the literature, rather vague descriptions can be found such as "An outlier can be defined to be an abnormal item among a group of otherwise similar items" (Choudhury and Das 1992, p. 92), to mention just one example. In such cases instead of using some distribution as the reference for outlyingness, some data adaptive approach is more promising. Again, the assumption is that the majority of the observations are generated by some target distribution, but the distribution itself is no longer specified in detail. To be able to identify outlying observations in such a context, an appropriate concept of distance is needed. Observations lying "far away" from the main body of the data are called outlying, where the "main body" as well as the notion of "far away" should be chosen from the data set itself.

The rest of this chapter is structured as follows. In Sect. 7.2, the task of identifying outliers when the target distribution is only vaguely determined is described in more detail. Sections 7.3, 7.4, and 7.5 describe three different approaches for tackling this problem. The chapter finishes with some concluding remarks.

7.2 The Identification of Outliers

The identification of outliers in a data set is an important task in data analysis. Often, graphical representations of the data can help in finding the unusual ones.

For multivariate (especially high-dimensional) data, such graphical representations are not easy to create. Hence, in this case analytical methods are needed.

7.2.1 Distance Based Outlier Identification

Among the various different approaches to multivariate outlier identification some popular methods are simultaneous one-step rules (Hawkins 1980; Davies and Gather 1993; Gather and Becker 1997) since they yield better results with respect to avoiding masking and swamping effects than e.g., sequential rules (so-called outlier testing, see Rosner 1975 for the fundamentals; also see Hawkins 1973; Hawkins 1980, p. 63ff.). Typically, these simultaneous outlier identification procedures use distance-based approaches.

Since often the target distribution is assumed to be the multivariate normal or at least some elliptically contoured or convex distribution, the distance concept used is that of Mahalanobis distance (or Euclidean distance based on a standardized version of the data), implying that location and scatter have to be estimated. And as this estimation has to be as less influenced by the outliers as possible, it should be done robustly.

A general concept for simultaneous outlier identification can be given as follows: consider a data set $\mathbf{X} \subset \mathbb{R}^p$, assuming that the majority of the N observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are generated by some elliptically contoured target distribution with location vector $\boldsymbol{\mu}$ and scatter matrix $\boldsymbol{\Sigma}$. Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ robustly by $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ and calculate the robust Mahalanobis distances $d_n = (\mathbf{x}_n - \widehat{\boldsymbol{\mu}})^T \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_n - \widehat{\boldsymbol{\mu}})$, $n = 1, \ldots, N$. Choose some appropriate critical value c and identify all observations $\mathbf{x}_n$ with $d_n > c$ as outliers. This rather general approach can be found in many sources (Becker and Gather 1999; Becker and Paris Scholz 2006; Hubert et al. 2008, and the literature cited therein, amongst many others), where it may be used in the raw version described above or with some refinements.

Although these methods work well in case of convex structures, the use of Mahalanobis type distances might be totally misleading if the shape of the data majority is of some different type. In other words, if we relax the assumptions on the target distribution, allowing for applications that are not normal and not even nearly normal at all, the established simultaneous outlier identification procedures may fail. For a simple example, consider the data set displayed in Fig. 7.1, mimicking the city arms of Halle (Saale) in Germany (see www.halle.de/de/Kultur-Tourismus/Stadtgeschichte/Wappen-der-Stadt-Halle/). The main part of the data follows the crescent-shaped form in the center, while the two point clouds lying above and below can be interpreted as outlying with respect to this not at all elliptical or convex shape. If we now apply one of the most commonly used simultaneous rules, which is based on the minimum covariance determinant (MCD) estimators (Rousseeuw 1985) in the reweighted version (Lopuhaä and Rousseeuw 1991; Hubert et al. 2008), the result is shown in Fig. 7.1. Here, points marked by a circle (○) stand for observations identified as outlying. Obviously, the procedure calls for the main body of

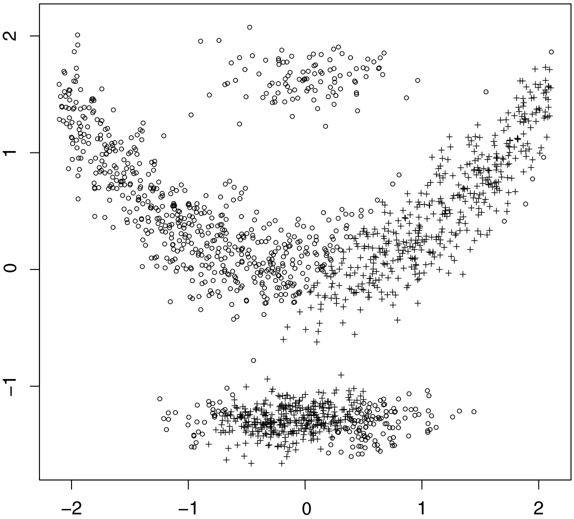


Fig. 7.1 Halle data: observations identified as outlying by the reweighted MCD approach, marked with ○

the data being elliptically contoured and cannot cope with the real shape. Hence, for situations like exemplified here, we call for new procedures not restricted to convex data structures.

7.2.2 The Main Body of the Data: Robust Subset Selection

Most of the common robust location and scatter estimation approaches assume that at least half of the data come from the target distribution, where at least the distribution class of this target is specified. From this assumption, some general approach to robust estimation has developed: in a first step, identify some small subset of the data which, at least with large confidence, consists only of observations stemming from the target distribution. Based on this outlier-free subset, an initial estimation is performed. This initial estimation is used to calculate each observation's distance as described above. The distances can then be used to decide whether the initial subset may be enlarged by further observations close enough to the main body of the data. If the initial choice of observations is enhanced, recalculate the estimators based on this enhanced subset. The reweighted MCD estimators mentioned above (Lopuhaä and Rousseeuw 1991; Hubert et al. 2008) are the probably most prominent example operating according to this approach.

The idea is appealing, hence, we propose to transfer it to the relaxed model assumptions: assume only that slightly more than 50 % of a data set forms some main structure of interest while the rest may be arbitrary observations. The general process is as follows:

- Determine a subset of the observations consisting only of points reflecting the main structure. If the only assumption is that this main structure is given by the majority of the observations, choose a subset of size 50 %.

- Based on this subset construct some appropriate distance measure to assess the observations' distances.
- If indicated by these distances enhance the initial subset to form the final subset.
- All observations not included in this final subset are declared to be outliers with respect to the main data structure.

Of course, the final subset can also be used to robustly estimate characteristics of the main data generating distribution. In the special case of elliptical or convex distributions, the approaches lead to robust estimates of location and covariance.

In the following Sections, different methods according to this general process are presented.

7.3 Flood Algorithm

The Flood Algorithm (Liebscher et al. 2012) is a proposal for robust estimation of multivariate location and scatter. It utilizes a two-stage approach where in a *first step* the initial data is projected into a two-dimensional space using self-organizing maps (Kohonen 1982) which are a specific kind of artificial neural networks. The self-organizing maps' algorithm preserves the topology of the underlying data while projecting it into the lower dimensional space (Kohonen 2001). Hence, one obtains a representation which still bears the majority of the information inherent in the initial data but at the same time is much easier to process. In particular, this dimensionality reduction allows for the visualization of the (projected) data in the form of the so-called U-landscape (Liebscher et al. 2012; Ultsch 1993). This landscape-like depiction gives some indication on the distance structure within the projected data and, therefore, within the initial data, too. In this plot small distances between the observations translate into valleys and basins while large distances translate into mountain ridges and plateaus. Thus, outliers—by definition featured by a large distance to the bulk of the data—can either be found on top of mountains (single outliers) or in small lakes/basins which are separated from the main basin by large mountain ridges (outlier clusters). In order to identify the aforementioned main basin corresponding to the outlier-free bulk of the data, the landscape is "flooded" in the *second step* of the Flood Algorithm. The flood level is raised until a basin is found which contains $h = \lfloor (N + p + 1)/2 \rfloor$ observations. These observations are subsequently used to estimate location and scatter by calculating the classical mean and covariance matrix of the h observations.

While the empirical results suggest that this approach gives robust estimates of location and scatter, outlier identification using robust Mahalanobis-type distances based on these estimates would suffer from the same deficiencies in the non-convex setting as outlined in Sect. 7.2.1. However, outlier identification may instead be done by using the distance information inherent in the aforementioned U-landscape as those give an indication on the *inter-point* distances and are therefore suitable to detect outliers in arbitrarily shaped data situations. To support this claim, we look at the Halle example again.

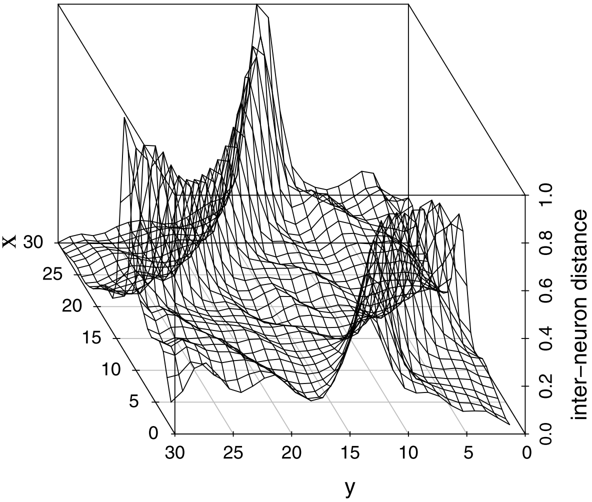
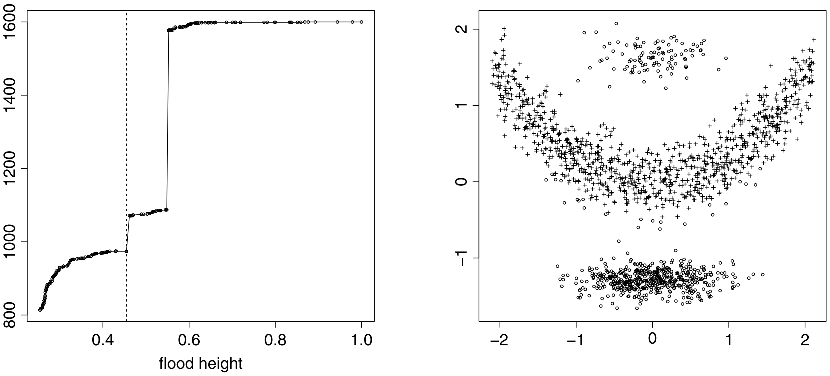


Fig. 7.2 U-landscape of the Halle example



(a) Flood-area-flood-height-curve of the Halle example

(b) Results obtained by the Flood Algorithm: non-outliers (+) and outliers (○)

Fig. 7.3 Results of the Flood Algorithm

Figure 7.2 shows the U-landscape for the Halle example. Clearly visible are two high mountain ridges which separate the landscape into three regions (i.e., three basins) where each one corresponds to one of the three clusters in the Halle data set. As the projection obtained by using self-organizing maps is non-linear, the U-shape of the main cluster (i.e., the middlemost basin) is no longer visible in the U-landscape. If the landscape in Fig. 7.2 is flooded and for each flood level the corresponding flooded area, measured by the number of observations in the largest basin, is noted, the curve shown in Fig. 7.3(a) results. We call this plot flood-area-flood-height-curve (Liebscher et al. 2012).

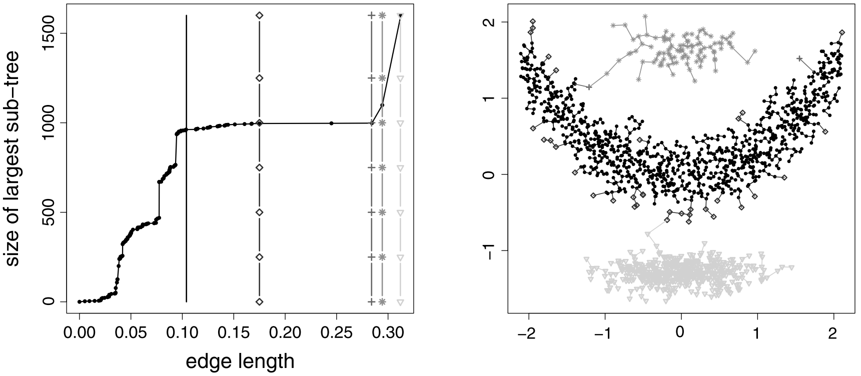
In the flood-area-flood-height-curve long line segments are of interest. They give an indication of single observations as well as clusters which are well separated from the surrounding observations. Long line segments in horizontal direction usually occur for single outliers, while vertical steps indicate the presence of clusters of outliers. With this knowledge in mind, we look for the smallest flood height associated with any long line segment. In our Halle example, this gives a flood height of 0.416 (dashed line in Fig. 7.3(a)). Applying this flood height to the Halle example results in the observations marked by $+$ in Fig. 7.3(b) being identified as the robust bulk of the data. The remaining observations, marked by $\circ$ in Fig. 7.3(b), are consequently flagged as outliers. While the overall separation between outliers and non-outliers is quite good, there are still some observations at the lower end of the U-shaped cluster which are not classified in a way one would expect them to be.

To summarize, the Flood Algorithm can be used to identify outlying observations, both single and clustered, in normal and non-normal data situations. While in the presented example the separation between outliers and non-outliers is not perfect because of the close proximity between the clusters, the overall performance is promising. Moreover, the flood-area-flood-height-curve provides an easy to interpret exploratory tool to decide on a suitable classification solution. On the other hand, the decision on the “appropriate” flood height is rather arbitrary and gives room for future research.

Computation of both the self-organizing map as well as the flooding procedure can be efficiently done, even for large numbers of observations and dimensions. The corresponding functions are available within the `R` environment (as part of the `restlos` package).

7.4 Pruned Minimum Spanning Tree

A simple method to measure distances is to use Euclidean distances and consider the data set $\mathbf{X}$ as a network. In such a network, a *node* reflects a particular observation and the length of an *edge* represents the distance (i.e., similarity) between two nodes. A complete network of all observations and all pairwise edges consists of N nodes and $N \cdot (N-1)/2$ edges. The set of edges is denoted by $\mathbf{E}$ and a particular edge $e \in \mathbf{E}$ is constituted by a pair of nodes, i.e., $e = \{\mathbf{x}, \mathbf{y}\}$. The weight $w(e)$ of an edge is given by the Euclidean distance between its constituting nodes, i.e., $w(e) = \|\mathbf{x} - \mathbf{y}\|_2$. Outliers are typically characterized by large distances to a major fraction of observations. In other words, such observations are separated from non-outlying observations. Hence, to decide on the separateness of an observations in a network, information about its incident edges have to be analyzed. However, considering a complete network complicates such an analysis since it contains all distance information including irrelevant ones. A sparser network, however, has to assure that all nodes (i.e., all observations) are contained and still all nodes are interconnected. A connection exists if a path, i.e., a set of pairwise adjacent edges, exists between any pair of nodes. Such a network is called a *spanning tree* if it contains exactly $N-1$ edges. If all edges have unique weight, there exists exactly one spanning tree whose length (the sum of all of its edges’ weights) is minimal.



(a) LC plot for Halle data set

(b) Halle data set with different subsets superimposed

Fig. 7.4 Results of pMST procedure for Halle data set

Such a spanning tree is called the *Minimum Spanning Tree* (MST) and is denoted by $\mathbf{T}^X = (\mathbf{X}, \mathbf{E}^*)$ with $|\mathbf{E}^*| = N - 1$. For more details on MSTs and their properties, see, e.g., Jungnickel (2008). A particular property is that for each node its nearest neighbour is directly connected by an edge. This property assures that local distance information from the complete network is preserved.

To decide which nodes are potential outliers, the MST is iteratively constructed. Define the following working sets: $\mathbf{P} = \varnothing$, $\mathbf{Q} = \varnothing$ and $\mathbf{S} = \mathbf{E}^*$. Next, the shortest edge $e_{\min} = \arg\min_{e \in \mathbf{S}} w(e)$ is omitted from $\mathbf{S}$ and stored in $\mathbf{P}$, the corresponding points are stored in $\mathbf{Q}$. If two or more edges have minimum length, one of these edges is chosen randomly. This step is repeated iteratively until $\mathbf{S}$ is empty and, hence, $\mathbf{P} = \mathbf{S}$. In $\mathbf{P}$ iteratively fragments of the MST are formed, where a fragment is defined as a connected subset of edges. Of particular interest is the size of the largest fragment of the MST found in $\mathbf{P}$. By taking a modified stopping criterion, this procedure can be used as a robust estimator (Kirschstein et al. 2013). An iterative pruning of the MST $\mathbf{T}^X$ by deleting the longest edge and retaining the larger subtree leads to the same result. This approach was first described by Bennett and Willemain (2001). The algorithm is called pMST procedure (pruned MST). Kirschstein et al. (2013) show that the robust estimator based on the pMST procedure achieves the maximum possible breakdown point (see the contribution by Müller, Chap. 5, for boundaries of breakdown points). Furthermore, simulation studies suggest robustness against various contamination schemes. The “constructing” approach presented here has the advantage that it basically equals the algorithm proposed by Kruskal (1956) to construct the MST based on a connected graph. In fact, the vector of MST fragment (or sub-tree) sizes can be seen as a by-product of Kruskal’s algorithm. Hence, the computational effort of the pMST procedure is $\mathcal{O}(N^2 \log N)$ (Jungnickel 2008).

As a graphical tool to decide which observations are outlying, the vector of MST fragment sizes is plotted against the corresponding vector of minimal edges’ lengths. This plot is called the length-connection (LC) plot, see Fig. 7.4(a) and Kirschstein et al. (2013). Single outliers are attached to an existing sub-tree by comparatively large edges, which manifest in the LC plot by long horizontal line segments. Clustered outliers are typically attached as a whole by a large edge, resulting in a long vertical line segment in the LC plot.

Since the pMST procedure uses local distance information to identify outliers, this method is particularly applicable for non-convex data. To illustrate, this feature the pMST procedure is applied to the Halle data set introduced above. Figure 7.4 shows the results of the pMST procedure. In Fig. 7.4(a), the LC plot for the Halle data set is depicted. Four remarkable breaks can be identified at edge lengths 0.104, 0.175, 0.284 and 0.294. At the leftmost break the general structure of the LC plot changes. Left of this marker the curve shows a rather steep ascent, whereas right of this marker the found sub-tree saturates by subsequent attachment of single observations. Up to the $\Diamond$-marker rather close observations are attached whereas the $+$-marker indicates the end of the saturation process by adding observations with clearly greater distance to the attached sub-tree. The already found sub-tree is depicted by the $\bullet$-points and black edges in Fig. 7.4(b), while by saturation the $\Diamond$- and $+$-marked points are added to this sub-tree. At the $+$-marker, this saturation ends and is followed by two vertical line segments indicating the attachment of the $*$-marked cluster of outliers depicted in Fig. 7.4(b). The final step indicates the attachment of the remaining lower cluster of outliers (marked by $\triangledown$ symbols) in Fig. 7.4(b).

To decide which observations should be regarded as outlying, a general advice is to look for structural breaks in the LC plot. In cases when a majority of data follows the same pattern, there is typically a steeply ascending beginning of the LC plot in the left part. When the slope flattens saturation is indicated which implies adding of scattered observations on the boundary of a cluster. A long vertical line segment marks a well-separated cluster being attached as a whole. A conservative choice is to handle all observations after a structural break (see the leftmost marker in Fig. 7.4(b)) as outliers. A more sensible approach is to look for comparatively long segments occurring after a structural break. In Fig. 7.4(a), this is indicated by the $\Diamond$-marker and results in a cluster of non-outlying observations with size 996. This implies that four non-outlying observations are mistakenly handled as outliers and no outlier is mistakenly identified as non-outlying. Even the conservative choice, indicated by the leftmost marker in Fig. 7.4(a) and the corresponding 962 $\bullet$-points in 7.4(b) does not declare outliers as non-outlying albeit failing to identify all non-outliers.

For all calculations, the `pMST` function from the `restlos` package is used with `R 2.15.1`.

7.5 RDELA Algorithm

The RDELA Algorithm (Liebscher et al. 2013) is another proposal for a robust estimator of multivariate location and scatter. It utilizes the Delaunay triangulation

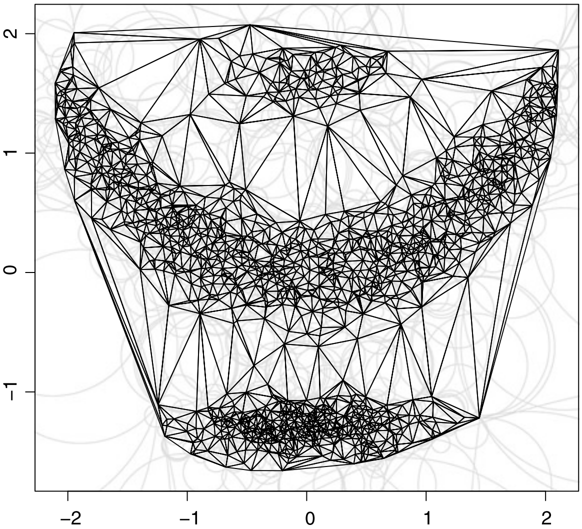


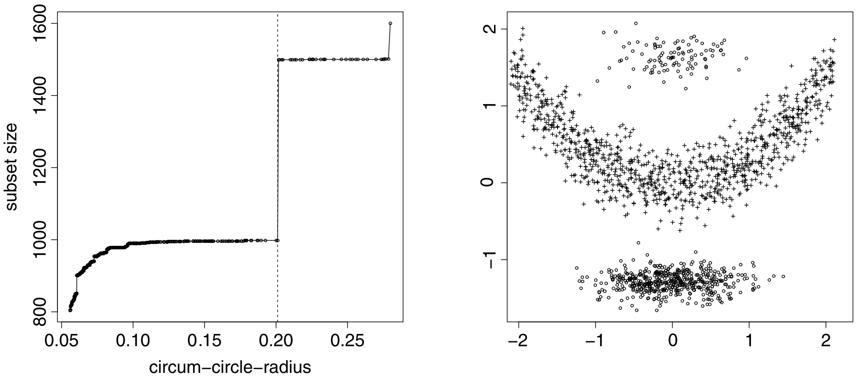
Fig. 7.5 Delaunay triangulation of the Halle example (with *circum-circles* in *grey*)

(Delaunay 1934) to identify the outlier-free bulk of data. Given a set of points $\mathbf{X}$ in $\mathbb{R}^p$ and the corresponding convex hull $CH(\mathbf{X})$, a triangulation is a decomposition of the polytope given by $CH(\mathbf{X})$ into simplices, i.e., into simple polytopes each consisting of exactly $p+1$ points of $\mathbf{X}$, see Fig. 7.5. The resulting simplices must not intersect each other and the original polytope's space must be entirely covered. In case of the Delaunay triangulation, it is additionally required that the decomposition is constructed in a way that the (hyper)spheres circumscribing each simplex do not contain any point of $\mathbf{X}$ (though the points of $\mathbf{X}$ are allowed to lie *on* the (hyper)spheres). If $\mathbf{X}$ is in general position, i.e. collinear and cocircular data configurations are not allowed, then the Delaunay triangulation is unique.

Following the calculation of the Delaunay triangulation the RDELA Algorithm selects some of the constructed simplices, where those selected are required to be neighboring and to have a small circum-sphere-radius. Selection is done iteratively. During the selection process the allowed radius is successively increased, resulting into an ever increasing number of simplices being taken into account. The vertices of the selected simplices constitute the identified subsample. Selection of the simplices continues until the size of the identified subsample equals $\lfloor (N+p+1)/2 \rfloor$. Location and scatter are finally estimated by calculating the mean and the covariance matrix of the observations in the subset. It is shown that this approach yields highly robust estimates (Liebscher et al. 2013).

As the radii give an indication of the (local) distances within the data set, in particular the radii of the simplices which have not yet been selected may be further analyzed in order to identify outliers.

Figure 7.5 shows the Delaunay triangulation of the Halle data set. While the triangles constructed within the clusters are rather small (with the corresponding circum-circle being small, too), the triangles constructed between the clusters are much bigger with larger radii. Applying the RDELA Algorithm on this triangulation



(a) Diagnostic plot from the RDELA Algorithm for the Halle example

(b) Results obtained by the RDELA Algorithm: non-outliers (+) and outliers (∘)

Fig. 7.6 Results of the RDELA Algorithm

and noting the radius and the corresponding size of the identified subsample in each step of the algorithm gives the curve shown in Fig. 7.6(a).

The diagnostic plot is interpreted in the same way as outlined for the Flood Algorithm or the pMST procedure. Therefore, we choose a circum-circle-radius of 0.201 (dashed line in Fig. 7.6(a)) and expand the initially found subset by attaching further simplices whose radius is smaller than 0.201. This approach yields the observations marked by + in Fig. 7.6(b). The remaining observations (marked by ∘ in Fig. 7.6(b)) are classified as outliers.

In this example, the RDELA Algorithm is able to perfectly separate between the outliers and non-outliers. General results (though not shown here) also suggest superior behaviour compared to the Flood Algorithm or the pMST procedure. However, calculation of the Delaunay triangulation is computationally much more expensive than calculation of the self-organizing map or the minimum spanning tree. While the number of observations does not pose a problem, the number of dimensions may become a limiting factor. This may also be a point for future research.

The RDELA function is available within the `restlos` package.

7.6 Conclusions

The tasks of robust estimation and outlier identification are closely related. Many of the commonly used robust estimators of location and scatter are based upon the selection of an outlier-free subset as a starting point for the estimation process. While these procedures are designed for the case of the non-outlying part of the data coming from some specified target distribution, the idea can also be transferred to a more general setting, where only rather mild assumptions are made on the data generating process. The three methods presented in this chapter follow this generalized approach. The results for the example data show that these methods can deal with situations where the main bulk of observations is in particular of non-elliptical shape, a setting which the most popular robust estimators like the MCD cannot cope with.

References

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: Wiley.

Becker, C., & Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association*, *94*, 947–955.

Becker, C., & Gather, U. (2001). The largest nonidentifiable outlier: a comparison of multivariate simultaneous outlier identification rules. *Computational Statistics & Data Analysis*, *36*, 119–127.

Becker, C., & Paris Scholz, S. (2006). Deepest points and least deep points: robustness and outliers with MZE. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, & W. Gaul (Eds.), *From data and information analysis to knowledge engineering* (pp. 254–261). Heidelberg: Springer.

Bennett, M., & Willemain, T. (2001). Resistant estimation of multivariate location using minimum spanning trees. *Journal of Statistical Computation and Simulation*, *69*, 19–40.

Choudhury, D. R., & Das, M. N. (1992). Use of combinatorics for unique detection of unknown number of outliers using group tests. *Sankhya. Series B*, *54*, 92–99.

Dang, X., & Serfling, R. (2010). Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties. *Journal of Statistical Planning and Inference*, *140*, 782–801.

Davies, P. L., & Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association*, *88*, 782–801.

Delaunay, B. (1934). Sur la sphere vide. *Izvestiâ Akademii Nauk SSSR. Otdelenie Tehničeskih Nauk*, *7*, 793–800.

Fieller, N. R. J. (1976). *Some problems related to the rejection of outlying observations*. Ph.D. Thesis, University of Hull, Hull.

Gather, U., & Becker, C. (1997). Outlier identification and robust methods. In G. S. Maddala & C. R. Rao (Eds.), *Handbook of statistics 15: robust inference* (pp. 123–143). Amsterdam: Elsevier.

Hampel, F. R., Rousseeuw, P. J., Ronchetti, E., & Stahel, W. (1986). *Robust statistics. The approach based on influence functions*. New York: Wiley.

Hawkins, D. M. (1973). Repeated testing for outliers. *Statistica Neerlandica*, *27*, 1–10.

Hawkins, D. M. (1980). *Identification of outliers*. London: Chapman & Hall.

Hubert, M., Rousseeuw, P. J., & van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science*, *23*, 92–119.

Jungnickel, D. (2008). *Graphs, networks and algorithms* (3rd ed.). Heidelberg: Springer.

Kirschstein, T., Liebscher, S., & Becker, C. (2013). *Robust estimation of location and scatter by pruning the minimum spanning tree*. Submitted for publication.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59–69.

Kohonen, T. (2001). *Self-organizing maps* (3rd ed.). Berlin: Springer.

Kruskal, J. (1956). On the shortest spanning subtree and the traveling salesman problem. *Proceedings of the American Mathematical Society*, *7*, 48–50.

Liebscher, S., Kirschstein, T., & Becker, C. (2012). The flood algorithm—a multivariate, self-organizing-map-based, robust location and covariance estimator. *Statistics and Computing*, *22*, 325–336. doi:10.1007/s11222-011-9250-3.

Liebscher, S., Kirschstein, T., & Becker, C. (2013). Rdela—a Delaunay-triangulation-based, location and covariance estimator with high breakdown point. *Statistics and Computing* doi:10.1007/s11222-012-9337-5.

Lopuhaä, H. P., & Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, *19*, 229–248.

Mara, W. (2011). *The Chernobyl disaster: legacy and impact on the future of nuclear energy*. New York: Marshall Cavendish.

Murphy, R. B. (1951). *On tests for outlying observations*. Ph.D. Thesis, Princeton University, Ann Arbor.

Pearson, E. S., & Chandra Sekar, C. (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, *28*, 308–320.

Rosner, B. (1975). On the detection of many outliers. *Technometrics*, *17*, 221–227.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In W. Grossman, G. Pflug, I. Vincze, & W. Wertz (Eds.), *Mathematical statistics and applications* (pp. 283–297). Dordrecht: Reidel.

Ultsch, A. (1993). Self-organizing neural networks for visualization and classification. In O. Opitz, B. Lausen, & R. Klar (Eds.), *Information and classification: concepts* (pp. 307–313). Berlin: Springer.

Chapter 8
Robustness for Compositional Data

Peter Filzmoser and Karel Hron

8.1 Introduction

Many real-world multivariate data sets are of compositional nature, which means that not the absolute reported information in variables but their ratios are informative. This situation frequently occurs in geochemistry, but also in biosciences or economics and many other applications (Pawlowsky-Glahn and Buccianti 2011). When analyzing a chemical composition of a rock, not the absolute values of the mass of the compounds (which depends on the size of the sample), but ratios provide a relevant picture of the multivariate data structure. Such observations, called in the following compositional data (or compositions for short), are popularly represented in proportions or percentages, i.e., as data with a constant sum constraint. This fact lead in the past to a mismatch of the concept of compositional data with that of constrained observations. In the latter case, natural requirements valid for compositional data are not met, like scale invariance (the information in a composition does not depend on the particular units in which the composition is expressed) and subcompositional coherence (information conveyed by a full composition should not be in contradiction with that coming from a sub-composition), see, e.g., Egozcue (2009) for details. The presented more general (and also a more natural) definition

P. Filzmoser (✉)
Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstrasse 8-10, Vienna, Austria
e-mail: P.Filzmoser@tuwien.ac.at

K. Hron
Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, 17. listopadu 12, Olomouc, Czech Republic
e-mail: hronk@seznam.cz

K. Hron
Department of Geoinformatics, Faculty of Science, Palacký University, tř. Svobody 26, Olomouc, Czech Republic

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_8, © Springer-Verlag Berlin Heidelberg 2013

was established in Aitchison (1986), where the logratio approach to the statistical analysis of compositional data was introduced. Since the specific properties of compositions naturally induce their own geometry (the Aitchison geometry) on the simplex, the sample space of compositions, the main effort is devoted to express the compositions in orthonormal coordinates, where the usual Euclidean rules already hold (Egozcue and Pawlowsky-Glahn 2006), and accommodate the standard statistical methods for their analysis.

Also in the case of compositional data, outlying observations can completely destroy results of a statistical analysis, comparing to those obtained from the homogeneous majority of observations in a data set. However, the specific geometric behavior of compositional data induces a different view of outliers compared to the usual case. For example, now obviously an observation with high absolute values on the compounds (parts) must not necessarily be an outlier, if the corresponding ratios between its parts follow the dominant data behavior. For this reason, not only the classical statistical methods, but even the robust ones cannot be applied directly to raw compositional data. This would lead (among other problems mentioned below) to a mismatch of regular and outlying observations.

In the following, a concise way how to perform multivariate statistical analysis of compositional data using the logratio approach will be presented. The next section introduces the Aitchison geometry on the simplex, together with the main inherent concepts of compositional data analysis and a way how to express (and interpret) the compositions in orthonormal coordinates. Section 8.3 shows how the classical and robust versions of standard multivariate methods like outlier detection, principal component analysis, correlation analysis and discriminant analysis can be applied to a compositional data set, together with interpreting the corresponding results. A real-world example on geochemical (compositional) data in presented in Sect. 8.4, and Sect. 8.5 concludes.

8.2 Geometrical Properties of Compositional Data

In addition to scale invariance and subcompositional coherence, another crucial property that characterizes compositional data is represented by their relative scale. The concept of relative scale naturally occurs already for most positive univariate data sets (Mateu-Figueras and Pawlowsky-Glahn 2008): although Euclidean distances within two pairs of samples taken at two rain gauges, {5; 10} and {100; 105} in mm are the same, quite probably, in the first case most observers would say there was double the total rain in the second gauge compared to the first, while in the second case they will say it rained a lot but approximately the same. Similarly, Euclidean distance between two multivariate observations does not reflect the relative growth in the compositional parts (concentrations of chemical elements, household expenditures on various commodities, etc.). Unfortunately, standard statistical methods completely ignore the relative scale concept since they rely on the usual Euclidean geometry in the real space (Eaton 1983). For this reason, the geometrical behavior of D-part compositional data $\mathbf{x} = \mathcal{C}(x_1, \dots, x_D)'$ is characterized by the Aitchison geometry, defined for compositions $\mathbf{x}$, $\mathbf{y}$ and a real constant a by the operations perturbation, $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \dots, x_D y_D)'$, powering, $a \odot \mathbf{x} = \mathcal{C}(x_1^a, \dots, x_D^a)'$ and the Aitchison inner product,

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$$

(note that the sums run over all log-ratios of the parts in the compositions $\mathbf{x}$ and $\mathbf{y}$), resulting in the Aitchison norm

$$\|\mathbf{x}\|_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a}$$

and the Aitchison distance,

$$d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_a.$$

The symbol $\mathcal{C}$ stands for the closure operation that rescales the resulting compositions to an arbitrarily chosen, but fixed constant sum constraint κ, like 100 in case of percentages, corresponding to the actual representation of the compositions in the D-part simplex,

$$\mathcal{S}^D = \left\{ (x_1, \dots, x_D)', \ x_i > 0, \ \sum_{i=1}^{D} x_i = \kappa \right\}.$$

Because all the relevant information in compositional data is contained in ratios between the parts, it is natural that zero compositional parts are not allowed for the analysis. According to the character of the occurrence of zeros, either as a result of an imprecise measurement of a trace element in the composition (i.e., rounding zeros) or the result of structural processes (structural zeros), special care has to be taken prior to a further processing of the observations (Aitchison and Kay 1999; Martín-Fernández et al. 2012).

The Aitchison geometry forms a Euclidean vector space of dimension $D - 1$, so it is possible to express the compositions in coordinates with respect to an orthonormal basis on the simplex, i.e., to obtain orthonormal coordinates of compositional data. The corresponding mapping $h : \mathcal{S}^D \rightarrow \mathbf{R}^{D-1}$, that results in the real vector $h(\mathbf{x}) = \mathbf{z} = (z_1, \dots, z_{D-1})'$, moves the Aitchison geometry to the standard Euclidean geometry in the real space (with vector addition, multiplication by a scalar and the Euclidean inner product), $h(\mathbf{x} \oplus \mathbf{y}) = h(\mathbf{x}) + h(\mathbf{y})$, $h(a \odot \mathbf{x}) = a \cdot h(\mathbf{x})$ and $\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle h(\mathbf{x}), h(\mathbf{y}) \rangle_e$. These properties are the reason why the mapping h is usually referred to as the isometric logratio (ilr) transformation (Egozcue et al. 2003).

Obviously, there are infinitely many possibilities how to choose the orthonormal basis on the simplex and construct the orthonormal coordinates. Unfortunately, there is no canonical basis on the simplex (D original compositional parts are represented by only $D - 1$ new coordinates), so that interpretable alternatives are needed. One possible choice represents the concept of balances (Egozcue and Pawlowsky-Glahn 2005), that enables an interpretation of the orthonormal coordinates in the sense of balances between groups of compositional parts. However, their construction usually assumes a prior knowledge of the studied problem. For this reason, we present below an "automated" version of balances, as described in Filzmoser et al. (2012a), that frequently occurs in different contexts of compositional data analysis (Hron et al. 2010, 2012). Explicitely, we obtain $(D-1)$-dimensional real vectors $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})', l = 1, \ldots, D$,

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \quad i = 1, \ldots, D-1, \tag{8.1}$$

where $(x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})$ stands for such a permutation of the parts of $\mathbf{x}$ that always the l-th compositional part fills the first position, $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)$. In such a configuration, the first ilr variable $z_1^{(l)}$ explains all the relative information (log-ratios) about the original compositional part x_l. The coordinates $z_2^{(l)}, \ldots, z_{D-1}^{(l)}$ then explain the remaining log-ratios in the composition (Fišerová and Hron 2011). Note that the only important position is that of $x_1^{(l)}$ (because it can be fully explained by $z_1^{(l)}$). The other parts can be chosen arbitrarily, because different ilr transformations are orthogonal rotations of each other (Egozcue et al. 2003). Of course, we cannot identify $z_1^{(l)}$ with the original compositional part x_l, but it explains all the information concerning x_l. Without loss of generality, we identify $z_i^{(1)} = z_i$ and use this simplified notation throughout the paper. Finally, the inverse ilr transformation, defined as $\mathbf{x} = \mathcal{C}(x_1, \ldots, x_D)' = h^{-1}(\mathbf{z})$, where

$$\begin{aligned}
x_1 &= \exp\left(\frac{\sqrt{D-1}}{\sqrt{D}} z_1\right), \\
x_i &= \exp\left(-\sum_{j=1}^{i-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j + \frac{\sqrt{D-i}}{\sqrt{D-i+1}} z_i\right), \\
&\quad i = 2, \ldots, D-1, \\
x_D &= \exp\left(-\sum_{j=1}^{D-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j\right),
\end{aligned} \tag{8.2}$$

is used to express the orthonormal coordinates back on the simplex.

For most statistical methods, an interpretation of the compositional data analysis in orthonormal coordinates is fully satisfactory. An exception is the biplot of principal component analysis which is related to the centered logratio (clr) transformation, resulting for a composition $\mathbf{x} = \mathcal{C}(x_1, \ldots, x_D)'$ in a real vector

$$\mathbf{y} = (y_1, \ldots, y_D)' = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)'.$$

Elements of $\mathbf{y}$ represent coefficients with respect to a generating system of compositions, i.e., the covariance matrix of a random composition $\mathbf{y}$ is positive semidefinite and thus the clr transformed data are not appropriate for a robust statistical analysis. Fortunately, there exists a linear relation between the clr coefficients and orthonormal coordinates (like those from (8.1)), $\mathbf{y} = \mathbf{V}\mathbf{z}$. The columns of the $D \times (D-1)$-matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{D-1})$ are formed by the clr transformation of the orthonormal basis vectors, resulting in coordinates $\mathbf{z}$, concretely

$$\mathbf{v}_{D-i} = \sqrt{\frac{i}{i+1}} \left(0, \ldots, 0, 1, -\frac{1}{i}, \ldots, -\frac{1}{i} \right)', \quad i = 1, \ldots, D-1.$$

Note that there is an important relation between the clr variables and the ilr transformation used in (8.1), namely $y_l = \sqrt{\frac{D}{D-1}} z_1^{(l)}$, $l = 1, \ldots, D$. This means that each clr variable captures all the relative information about the compositional part x_l, and y_l is proportional to $z_1^{(l)}$.

Although measures of location and variability of a random composition can even be expressed directly on the simplex using the mean value of the Aitchison distance as the center and the total variance,

$$\mathrm{cen}(\mathbf{x}) = \operatorname{argmin}_{\boldsymbol{\eta}} \mathsf{E}\big[d_a^2(\mathbf{x}, \boldsymbol{\eta})\big], \qquad \mathrm{totvar}(\mathbf{x}) = \mathsf{E}\big[d_a^2\big(\mathbf{x}, \mathrm{cen}(\mathbf{x})\big)\big],$$

respectively (Pawlowsky-Glahn and Egozcue 2002), it is usually preferred to capture location and variability of compositions directly in coordinates using the expectation $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The variances and covariances of coordinates can be expressed using variances of log-ratios of parts of the original composition, $\mathrm{var}(\ln \frac{x_i}{x_j})$, $i, j = 1 \ldots, D$, that are advantageous for interpretation purposes (Fišerová and Hron 2011). If a sample $\mathbf{z}_1, \ldots, \mathbf{z}_n$ is given for the coordinate $\mathbf{z}$, one usually arrives at the arithmetic mean $\overline{\mathbf{z}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i$ and the sample covariance matrix $\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{z}_i - \overline{\mathbf{z}})(\mathbf{z}_i - \overline{\mathbf{z}})'$. Clearly, these characteristics have a breakdown point of zero, and thus they are highly sensitive to outlying observations, see the contribution by Müller, Chap. 5. Consequently, proper robust alternatives are needed, see the contribution by Rousseeuw and Hubert, Chap. 4. Because of different representations of compositions in coordinates, the affine equivariance of the corresponding estimators is crucial. In the following, we use the Minimum Covariance Determinant (MCD) estimator (Maronna et al. 2006), which is advantageous in particular for computational reasons (Rousseeuw and Van Driessen 1999). Using the relations

$$\mathrm{cen}(\mathbf{x}) = h^{-1}\big(\mathsf{E}\big[h(\mathbf{x})\big]\big), \qquad \mathrm{totvar}(\mathbf{x}) = \mathrm{trace}(\boldsymbol{\Sigma}),$$

the MCD algorithm can be directly used to estimate the center and the total variance of compositional data.

8.3 Multivariate Statistical Methods for Compositional Data

The scope of the statistical analysis of compositions does not differ from the case of standard multivariate data. Starting with an unsupervised case, the analyst is interested in identifying groups and main patterns of the data as well as in detecting outliers that depart from the main data cloud. Here, principal component analysis and outlier detection provide a first insight into the multivariate data structure. In the second step, natural groups of observations and their variables are analyzed in a supervised manner using discriminant analysis and correlation analysis, respectively, although the latter one can also be advantageously applied to support the unsupervised case. The main aspects and interpretation of the above mentioned methods are described in the following sections. Since all the methods strongly rely on a proper (robust) estimation of location and covariance, special care will be devoted to the compositional data specificity also in this context.

8.3.1 Outlier Detection

Outlier detection belongs to starting points of each exploratory data analysis, as outliers give valuable information on data quality, and they are indicative of atypical phenomena. Outlier detection methods usually assume a certain type of (theoretical) distribution of the main data cloud; in case of an elliptical distribution, the corresponding criterion can be based on the Mahalanobis distance, defined for a sample of compositions in coordinates as

$$\mathrm{MD}(\mathbf{z}_i) = \left[(\mathbf{z}_i - \mathbf{t})'\mathbf{C}^{-1}(\mathbf{z}_i - \mathbf{t})\right]^{1/2}, \quad i = 1, \ldots, n. \tag{8.3}$$

Here, $\mathbf{t} = \mathbf{t}(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ and $\mathbf{C} = \mathbf{C}(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ are location and covariance estimators, respectively. The choice of the estimators is crucial for the quality of multivariate outlier detection. Taking the classical estimators arithmetic mean and sample covariance matrix often leads to useless results, because these estimators themselves are influenced by deviating data points. For this reason, robust counterparts, like the above mentioned MCD estimator, need to be taken that downweight the influence of outliers on the resulting location and covariance estimation statistics, see the contributions by Mosler (Chap. 2) and Rousseeuw and Hubert (Chap. 4). Under the assumption of multivariate normal distribution on the simplex, i.e., normal distribution of the orthonormal coordinates (Mateu-Figueras and Pawlowsky-Glahn 2008), the (classical) squared Mahalanobis distances follow a chi-square distribution with $D-1$ degrees of freedom, see, e.g., Maronna et al. (2006). This distribution might also be considered for the robust case, and a quantile, e.g., 0.975, can be used as a cut-off value separating regular observations from outliers. In case of compositional data, the values of the Mahalanobis distances do not depend on the chosen coordinates, i.e., also $\mathrm{MD}(\mathbf{z}_i) = \mathrm{MD}(\mathbf{z}_i^{(l)})$, $l = 2, \ldots, D$, if affine equivariant estimators of location and covariance are used (Filzmoser and Hron 2008).

A more advanced approach for the cut-off value was used in Filzmoser et al. (2012a). This accounts for the actual numbers of observations and variables in the data set, and it tries to distinguish among extremes of the data distribution and outliers coming from a different distribution. The paper also provides an overview of graphical methods for an interpretation of the multivariate outliers, available in the R package `mvoutlier`. In particular, the coordinates $z_1^{(l)}$ can be advantageously used for univariate presentations in order to reveal outliers, connected with one or more concrete compositional parts (see Sect. 8.4 for an example).

8.3.2 Principal Component Analysis and the Compositional Biplot

Also for the well-known method principal component analysis (PCA), the proper estimation of location ($\mathbf{t}$) and covariance ($\mathbf{C}$) plays an important role. Let $\mathbf{C} = \mathbf{G_z L G_z'}$ denote a spectral decomposition of the estimated covariance matrix $\mathbf{C}$, with the diagonal matrix $\mathbf{L}$ of eigenvalues and the matrix $\mathbf{G_z}$ of eigenvectors of $\mathbf{C}$. Then PCA results in a linear transformation

$$\mathbf{z}_i^* = \mathbf{G}_\mathbf{z}'(\mathbf{z}_i - \mathbf{t}), \tag{8.4}$$

of the coordinates into new variables (principal components) such that the first principal component has the largest possible variance (accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. Although both scores $\mathbf{z}_i^*$ and loadings (columns of the matrix $\mathbf{G_z}$) of the principal components could also be interpreted directly in orthonormal coordinates, it is rather common to transform the loadings back to the clr space, $\mathbf{G_y} = \mathbf{V G_z}$, where the affine equivariance property of the MCD estimator was utilized (Filzmoser et al. 2009). The scores in the clr space are identical to the scores of the ilr space, except that the additional last column of the clr score matrix has entries of zero. Finally, the transformed loadings and scores are used to construct the biplot of clr transformed compositional data (Aitchison and Greenacre (2002), also called “compositional biplot”). Although the purpose of the compositional biplot is the same as for the standard one (Gabriel 1971), i.e., to provide a planar graph that represents a rank-two approximation of both the observations (PCA scores, plotted as points) and variables (loadings, rays) of multivariate data, its interpretation is different: The main interest is in the links (distances between vertices of the rays); concretely, for the rays i and j ($i, j = 1, \ldots, D$) the link approximates the log-ratio variance $\mathrm{var}(\ln \frac{x_i}{x_j})$. Hence, when the vertices coincide, or nearly so, then the ratio between x_i and x_j is constant, or nearly so. Consequently, this characteristic replaces the thinking in terms of correlation coefficients between two coordinates (standard variables). In addition, directions of the rays signalize where observations with dominance of the corresponding compositional part are located.

8.3.3 Correlation Analysis

Correlation analysis is not applicable to the original compositional parts. An exception is the case where compositional data (or constrained data) are modeled with a Dirichlet distribution. There, correlation analysis it is still a popular choice for expressing the strength of a linear relation between the parts of a positive vector with constant sum constraint. However, the interpretation of the resulting correlation coefficients is misleading due to negative bias of the covariance structure of constrained observations (Aitchison 1986). In fact, a correlation analysis of compositions is meaningful only if it is applied to orthonormal coordinates. Here the concept of balances provides a possibility of interpretable results, when the groups of compositional parts are clearly stated, like based on the results of the compositional biplot. However, practical experiences show that also in such cases a deeper insight to the studied problem is usually required (Filzmoser and Hron 2009).

An "unsupervised version" of correlation analysis is given by the multiple correlation coefficient applied to the special choice of balances according to (8.1). The multiple correlation coefficient ϱ_l between a balance $z_1^{(l)}$ and the remaining balances $\mathbf{z}_l = (z_2^{(l)}, \ldots, z_{D-1}^{(l)})'$ can be expressed as

$$\varrho_l^2 = 1 - \frac{|\boldsymbol{\Sigma}|}{\mathrm{var}(z_1^{(l)}) \cdot |\boldsymbol{\Sigma}_l|}, \quad l = 1, \ldots, D, \tag{8.5}$$

where $\boldsymbol{\Sigma}_l$ denotes the covariance matrix of $\mathbf{z}_l$ (Johnson and Wichern 2007). The estimation of the theoretical characteristics can again be performed using the MCD estimator. Note that any rotation of $\mathbf{z}_l$ (another choice of the corresponding balances) would lead to the same value of the coefficient ϱ_l (Filzmoser and Hron 2009). The interpretation of the multiple correlation coefficient for the above chosen coordinates can be directly derived from the standard case and the properties of balances. As the balance $z_1^{(l)}$ explains all the relative information (log-ratios) about the compositional part x_l and the vector $\mathbf{z}_l$ expresses orthonormal coordinates for the subcomposition $\mathcal{C}(x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})'$, the coefficient $\varrho_l \in [0, 1]$ can be interpreted as a measure of strength of the linear relation between relative information on x_l and the rest of the composition. Consequently, small values of ϱ_l signalize an exceptional behavior of the part x_l with respect to the other compositional parts. In addition, a big difference between the classical and robust version of the coefficient indicates that the possible relation is driven by outliers.

8.3.4 Discriminant Analysis

Discriminant analysis is a tool for supervised classification. It can be directly applied after any orthonormal coordinates are chosen to express the compositions in the real space (Filzmoser et al. 2012b). This holds for both well established rules, the Bayesian and the Fisher discriminant rule (Johnson and Wichern 2007). For the purpose of exploratory analysis, the Fisher rule seems more convenient. Having g groups of observations with p_j as their prior probabilities ($\sum_{j=1}^{g} p_j = 1$), their expectations $\boldsymbol{\mu}_j$ and the covariance matrices $\boldsymbol{\Sigma}_j$, $j = 1, \ldots, g$, the Fisher rule is based on the matrix $\mathbf{B}$ describing the variation between the groups, and the matrix $\mathbf{W}$ that stands for the within groups covariance matrix. If the notation $\boldsymbol{\mu} = \sum_{j=1}^{g} p_j \boldsymbol{\mu}_j$ for the overall weighted mean of all populations is used, the two matrices are defined as

$$\mathbf{B} = \sum_{j=1}^{g} p_j(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})', \qquad \mathbf{W} = \sum_{j=1}^{g} \boldsymbol{\Sigma}_j.$$

Under the assumption of equal group covariance matrices, it can be shown that the best separation of the group means can be achieved by maximizing

$$\frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}} \quad \text{for } \mathbf{a} \in \mathbf{R}^{D-1}, \ \mathbf{a} \neq \mathbf{0}, \tag{8.6}$$

in the sample version preferably by using robust estimates of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$. The solution of this maximization problem is given by the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_l$ of the matrix $\mathbf{W}^{-1}\mathbf{B}$, scaled so that $\mathbf{v}_i'\mathbf{W}\mathbf{v}_i = 1$ for $i = 1, \ldots, l$. The number l of strictly positive eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ turns out to be $l \leq \min(g-1, D-1)$. The main advantage of the Fisher discriminant rule is its ability for dimension reduction. Concretely, by projecting the data in the space of the first two eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$, one obtains a data presentation in the plane that best captures the differences among the groups. Note that the resulting scores of these projections are invariant to the choice of the orthonormal coordinates, see Filzmoser et al. (2012b) for details.

8.4 Example

The methods described above are illustrated with a data set from geochemistry, the so-called *Oslo transect data*, see Reimann et al. (2007), and references therein. The data set is available in the R package `rrcov` as data `OsloTransect` (R Development Core Team 2012). Samples of different plant species (birch, mountain ash, fern and spruce) were collected along a 120 km transect running through Oslo (Norway), and the concentration of 25 chemical elements for the sample materials are reported. Since different parts were taken from the four sample materials (like leaves and bark), in total nine groups are available, where each group consists of 40 samples.

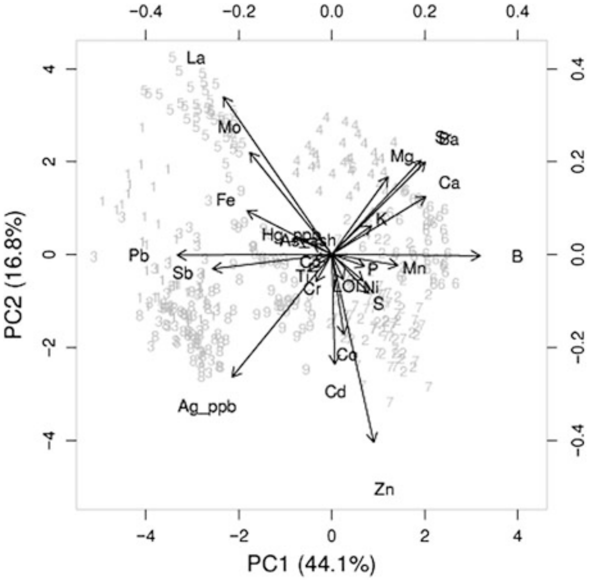


Fig. 8.1 Biplot of the first two robust principal components for all observations and variables. The numbers correspond to the data groups: birch bark (*1*), birch leaves (*2*), birch wood (*3*), fern (*4*), moss (**5**), European mountain ash leaves (*6*), spruce needles (**7**), spruce tree wood (*8*), spruce tree twigs (*9*)

The goal of this section is not a complete analysis of the *Oslo transect data*, but rather an illustration of the robust multivariate methods applied to subsets of this compositional data set. Concentrations of chemical elements are indeed compositions, because an increase of one part will automatically result in a decrease of the other parts, and thus ratios rather than absolute concentrations are of interest.

Figure 8.1 shows a compositional biplot of a robust PCA carried out with the complete data set. The first two principal components explain about 60 % of the total variability. Several variable groups can be identified, pointing at different processes in the plants. In addition, the data groups are well visible as score groups in the plot (each group is represented as an own number). Consequently, some groups show a higher dominance of certain variables, like a dominance of boron (B) in group 6 (European Mount Ash leaves).

Multivariate outlier detection is illustrated with the variable set Ba, Ca, Cu, Mn, Pb, Sr, Zn and the data group birch bark (the samples are now numbered from 1 to 40). In addition, 5 samples from the group birch wood (numbered as 41–45) are taken. Outlier detection should be helpful to identify the 5 samples from the different sample material. Indeed, applying the function `mvoutlier.CoDa` from the R package `mvoutlier` reveals the observations 5, 9, 25, 29, 40, 41, 42, 43, 44 and 45 as multivariate outliers. Several diagnostic plots are provided for finding an interpretation of the outliers. Figure 8.2 shows a parallel coordinate plot with the outliers shown by black lines. The presented axes are the ilr variables constructed with (8.1), and so they represent all the relative information of the corresponding chemical elements. It can be seen that the outliers have a very different behavior than the regular observations. For example, they are much less dominated by Cu and Mn.

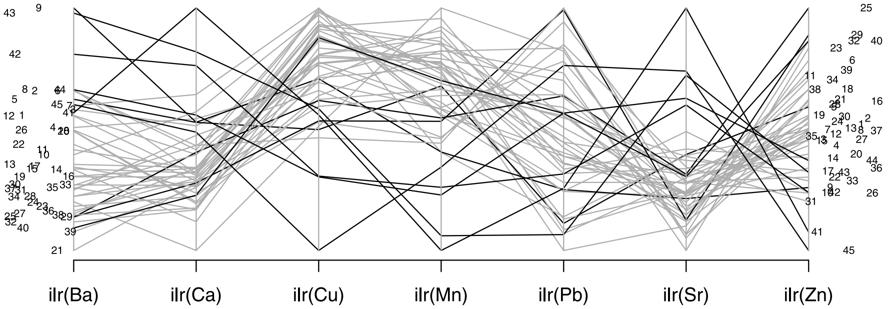


Fig. 8.2 Parallel coordinate plot for the ilr variables: multivariate outliers are shown by *black lines*; the numbers are the sample numbers: 1–40 for birch bark samples, and 41–45 for birch wood samples

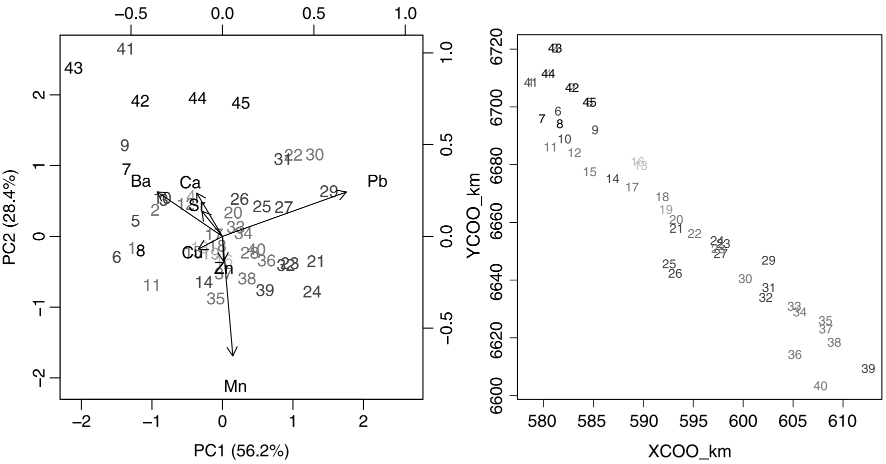


Fig. 8.3 Biplot (*left*) and plot of the spatial coordinates (*right*) of the data subset; the darker the symbols for the samples, the higher is their "degree of outlyingness"; the birch bark samples are numbered from 1–40, the birch wood samples from 41–45

Further plots for interpreting the multivariate outliers are shown in Figure 8.3. The biplot (left) reveals the group 41–45 as dominated by Ba and Ca, and with a shortage of Mn, compared to the other observations. The grey scale of the symbols corresponds to the "level of outlyingness", and it is also used in the further plots. On the right-hand side we see the samples plotted at their spatial coordinates. The birch wood samples originate from the North–West part of the region, and they are taken from the same locations as samples 1–5 (the symbols are over-plotted). Obviously, the sample material has more effect on the chemical composition than the origin of the samples.

The biplot in Figure 8.3(left) suggests that the variables are not highly related to each other. For example, the element Pb seems to be quite different from the other

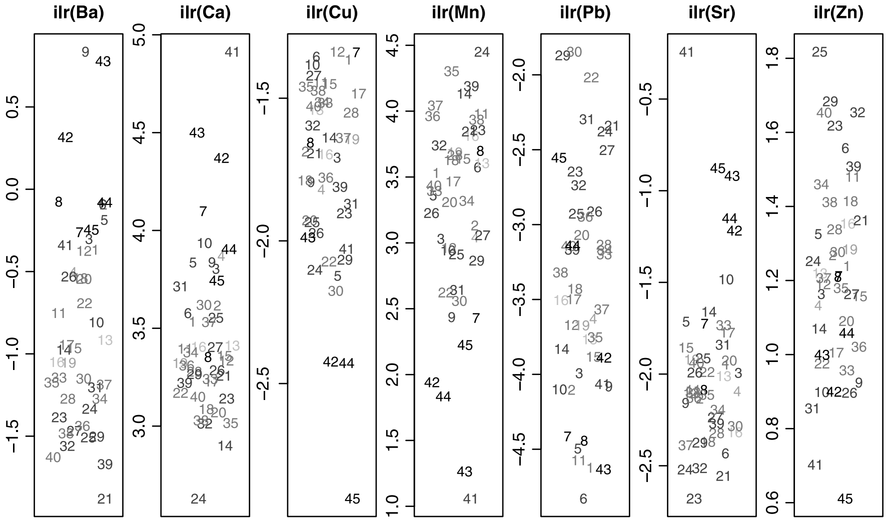


Fig. 8.4 Univariate scatter plots for the single ilr variables, including all the relative information of the corresponding chemical element; the outlier group 41–45 has a quite different chemical composition

elements. Computing usual correlations would not be appropriate here, but it is possible to compute the squared multiple correlation between Pb and the remaining elements using (8.5). The resulting robust (MCD-based) squared multiple correlation coefficient between the balance representing Pb and the remaining balances is 0.82, so not particularly high. In comparison, using classical estimators leads to a coefficient of 0.65, which reflects the influence of the outliers.

In a final outlier diagnostic in form of univariate scatter plots in Figure 8.4, we show again the ilr coordinates including all the relative information of an element, similar to Figure 8.2. This presentation can be used to see in detail the composition of the samples. The group 41–45 of birch wood samples is more dominant in Ba, Ca, and Sr, and much less dominant in Cu and Mn. Also some of the remaining outliers 5, 9, 25, 29, 40 can be found on extreme positions in this plot, justifying the reason for their outlyingness.

Robust discriminant analysis is applied using all observations of the *Oslo transect* data set, but only the variables Ba, Ca, Cu, Mn, Pb, Sr, Zn (they have no problem with rounding artefacts). Figure 8.5(left) shows the projection of the data onto the space of the first two Fisher discriminant directions. The different data groups (shown by different symbols) are quite well separated. However, single outliers are visible, and thus a robust analysis is useful. Figure 8.5(right) is an attempt to learn about the discrimination ability of the established discriminant rules. Here we use 5-fold cross-validation: The data set is randomly split into five parts of about equal size, and in each step 4 parts are used as training set and the remaining part as validation set. The discriminant rule is established from the training data and it is evaluated for the validation data. This is done until each of the five parts once had the

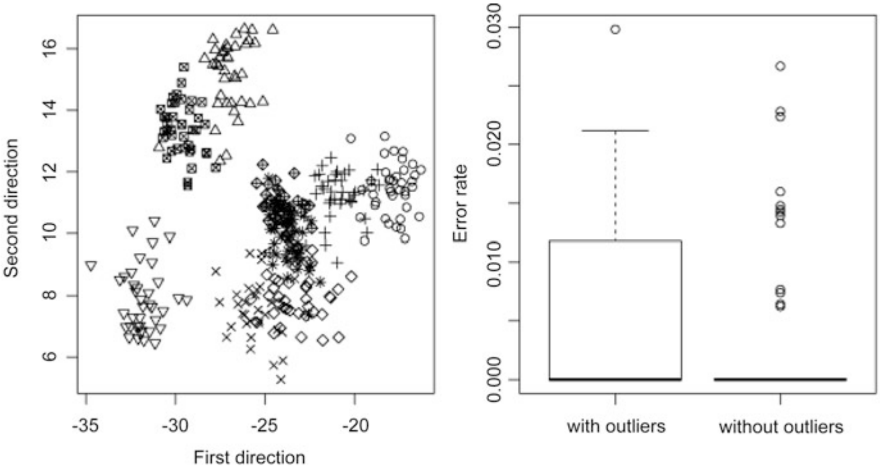


Fig. 8.5 Robust discriminant analysis using all samples (nine groups) and selected variables; the groups are well visible in the projection on the first two Fisher directions (*left*); the discriminant rules are evaluated with 5-fold cross-validation (*right*), by using all data in the validation sets (*left boxplot*) and removing the outliers from the validation sets (*right boxplot*)

role of the validation set. In each trial, we calculate the proportion of misclassified observations. In order to have a more general picture, the whole process is repeated 100 times, each time with new random assignments for 5-fold cross-validation. The left boxplot in the figure shows the resulting proportions of misclassification (error rate). Robust discriminant analysis is very effective here: the median error rate is zero, and at most 3 % error rate were reached. For generating the right boxplot we remove the outlying observations from the validation sets (they are still included in the training sets) by applying outlier identification in advance to each group separately. This gives a clearer picture of a robust discrimination rule, because outliers are not aimed to be classified correctly, in contrast to regular observations. The boxplot reveals that nearly all regular observations are correctly classified, and that the outliers in the training sets have no effect on the quality of the discrimination rules.

8.5 Conclusions

Compositional data, resulting from many real-world phenomena, require special transformations before the standard statistical tools can be used. Their characteristic is the fact that ratios between the variables carry the relevant information, and not the variable values directly. The suggested ilr transformation of (8.1) not only allows to analyze this relative information, but it is also helpful for the interpretation. Moreover, the resulting coordinates avoid the singularity problem of the clr transformation, and thus multivariate estimators like MCD can be applied. Since the MCD estimator is affine equivariant, the results from a robust multivariate analysis are not depending on the choice of the specific ilr transformation.

Tools for classical and robust compositional data analysis are available for example in the R packages `compositions`, `robCompositions`, and `mvoutlier`.

Acknowledgements The authors gratefully acknowledge the support by the Operational Program Education for Competitiveness—European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic).

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall.

Aitchison, J., & Greenacre, M. (2002). Biplots of compositional data. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *51*, 375–392.

Aitchison, J., & Kay, J. W. (1999). *Possible solutions of some essential zero problems in compositional data analysis*. Available at http://dugi-doc.udg.edu/bitstream/10256/652/1/Aitchison_Kay.pdf. Cited in March 12, 2012.

Eaton, M. (1983). *Multivariate statistics. A vector space approach*. New York: Wiley.

Egozcue, J. J. (2009). Reply to "On the Harker Variation Diagrams;..." by J.A. Cortés. *Mathematical Geosciences*, *41*, 829–834.

Egozcue, J. J., & Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, *37*, 795–828.

Egozcue, J. J., & Pawlowsky-Glahn, V. (2006). Simplicial geometry for compositional data. In A. Buccianti, G. Mateu-Figueras, & V. Pawlowsky-Glahn (Eds.), *Compositional data in the geosciences: from theory to practice* (pp. 145–160). London: Geological Society.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, *35*, 279–300.

Filzmoser, P., & Hron, K. (2008). Outlier detection for compositional data using robust methods. *Mathematical Geosciences*, *40*, 233–248.

Filzmoser, P., & Hron, K. (2009). Correlation analysis for compositional data. *Mathematical Geosciences*, *41*, 905–919.

Filzmoser, P., Hron, K., & Reimann, C. (2009). Principal component analysis for compositional data with outliers. *Environmetrics*, *20*, 621–635.

Filzmoser, P., Hron, K., & Reimann, C. (2012a). Interpretation of multivariate outliers for compositional data. *Computational Geosciences*, *39*, 77–85.

Filzmoser, P., Hron, K., & Templ, M. (2012b). Discriminant analysis for compositional data and robust parameter estimation. *Computational Statistics* doi:10.1007/s00180-011-0279-8.

Fišerová, E., & Hron, K. (2011). On interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences*, *43*, 455–468.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*, 453–467.

Hron, K., Filzmoser, P., & Thompson, K. (2012). Linear regression with compositional explanatory variables. *Journal of Applied Statistics*, *39*, 1115–1128.

Hron, K., Templ, M., & Filzmoser, P. (2010). Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics & Data Analysis*, *54*, 3095–3107.

Johnson, R., & Wichern, D. (2007). *Applied multivariate statistical analysis* (6th ed.). London: Prentice-Hall.

Maronna, R., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: theory and methods*. New York: Wiley.

Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P., & Palarea-Albaladejo, J. (2012). Model-based replacement of rounded zeros in compositional data: classical and robust approaches. *Computational Statistics & Data Analysis*, *56*, 2688–2704.

Mateu-Figueras, G., & Pawlowsky-Glahn, V. (2008). A critical approach to probability laws in geochemistry. *Mathematical Geosciences*, *40*, 489–502.

Pawlowsky-Glahn, V., & Buccianti, A. (2011). *Compositional data analysis: theory and applications*. Chichester: Wiley.

Pawlowsky-Glahn, V., & Egozcue, J. J. (2002). BLU estimators and compositional data. *Mathematical Geology*, *34*, 259–274.

R Development Core Team (2012). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reimann, C., Arnoldussen, A., Boyd, R., Finne, T. E., Koller, F., Nordgullen, O., & Englmair, P. (2007). Element contents in leaves of four plant species (birch, mountain ash, fern and spruce) along anthropogenic and geogenic concentration gradients. *Science of the Total Environment*, *377*, 416–433.

Rousseeuw, P., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*, 212–223.

Part II
Regression and Time Series Analysis

Chapter 9
Least Squares Estimation in High Dimensional Sparse Heteroscedastic Models

Holger Dette and Jens Wagener

9.1 Introduction

Consider a linear regression model of the form

$$Y_i = x_i^T \beta_0 + \varepsilon_i, \quad i = 1, \dots, n, \tag{9.1}$$

where $Y_i \in \mathbb{R}$, x_i is a p-dimensional covariate, $\beta_0 = (\beta_{0,1}, \dots, \beta_{0,p})^T$ the unknown vector of parameters and the ε_i are i.i.d. random variables. We assume that p is fixed and most of the components of the parameter vector β_0 vanish. This scenario is called sparse linear regression model in the literature. The problem of identifying significant components of the vector x and estimating the corresponding components of the parameter β_0 has found considerable interest in the last 20 years. A very well-known tool are penalized least squares methods because they provide an attractive methodology to select variables and estimate parameters in sparse linear models. We mention bridge regression (Frank and Friedman 1993), Lasso (Tibshirani 1996), the nonnegative Garotte (Breiman 1995), SCAD (Fan and Li 2001), least angle regression (Efron et al. 2004), the elastic net (Zou and Hastie 2005), the adaptive Lasso (Zou 2006), the Dantzig selector (Candes and Tao 2007) and Bayesian Lasso (see the contribution by Konrath, Fahrmeir and Kneib, Chap. 10) among others. Some of these meanwhile "classical" concepts are briefly explained in Sect. 9.2.

Much of the aforementioned literature concentrates on the case of linear models with independent identical distributed errors. To our best knowledge, there has been no attempt to investigate bridge estimators and the adaptive Lasso in high dimensional linear models with heteroscedastic errors.

H. Dette (✉) · J. Wagener
Fakultät für Mathematik, Ruhr-Universität Bochum, 44780 Bochum, Germany
e-mail: holger.dette@rub.de

J. Wagener
e-mail: jens.wagener@rub.de

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_9, © Springer-Verlag Berlin Heidelberg 2013

The purpose of this chapter is to present some of the commonly used methods for variable selection in high dimensional sparse linear regression methods with a special focus on heteroscedastic models. We give a careful explanation on how the choice of the regularizing parameter affects the quality of the statistical inference (such as conservative or consistent model selection). In Sect. 9.3, it is demonstrated that (under suitable assumption on the regularizing parameter) the model selection properties of the commonly used estimators in sparse linear regression models still persist under heteroscedasticity. However, the bridge estimators and the adaptive Lasso estimators of the important parameters are asymptotically normally distributed with the optimal rate, but with a suboptimal variance. In this section we also present an approach introduced in Wagener and Dette (2012) which is especially designed to address heteroscedasticity. For the case where the dimension p is fixed and the sample size n is increasing, we present some results regarding consistency and asymptotic normality. In particular, the new methods possess an oracle property in the sense of Fan and Li (2001), which means that they perform consistent model selection (see Sect. 9.3 for a precise definition) and estimate the nonvanishing parameters with the optimal rate and variance. The final Sect. 9.4 is devoted to a small numerical study, which investigates some of the theoretical properties for realistic sample sizes.

9.2 Penalized Least Squares Estimators

9.2.1 Bridge Regression

Frank and Friedman (1993) introduced the so called ‘bridge regression’, which is defined as the minimizer of an objective function penalized by the L^q-norm ($q > 0$), that is

$$L(\beta) = \sum_{i=1}^{n} \big(Y_i - x_i^T \beta\big)^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|^q, \tag{9.2}$$

where λ_n is a regularization parameter which converges to infinity with increasing sample size. The procedure shrinks the estimates of the parameters in the model (9.1) towards 0. Throughout this paper, we will always assume that $0 < q \leq 1$. In this case, it was shown by Knight and Fu (2000) that this procedure identifies the non vanishing components of the parameter β with positive probability and that the corresponding estimators are asympotically normal distributed (after an appropriate standardization). However, the corresponding optimization problem is not convex if $0 < q < 1$ due to the non-convex penalty. Therefore, the problem of determining the estimator minimizing (9.2) is a challenging one. Standard tools of convex optimization theory are not applicable and multiple local minima might exist.

9.2.2 Lasso and Adaptive Lasso

The celebrated Lasso (Tibshirani 1996) corresponds to a bridge estimator with $q = 1$. The asymptotic behavior of these estimators was investigated by Knight and Fu (2000), who established asymptotic normality of the estimators of the non-zero components of the parameter vector and showed that the estimator sets some parameters exactly to 0. Thereby the Lasso performs model selection and parameter estimation in a single step with computational cost growing polynomially with the sample size. Fan and Li (2001) argued that any reasonable estimator should be unbiased, continuous in the data, should estimate zero parameters as exactly zero with probability converging to one (consistency for model selection) and should have the same asymptotic variance as the ideal estimator in the correct model. They called this the 'oracle property' of an estimator, because such an estimator is asymptotically (point-wise) as efficient as an estimator which is assisted by a model selection oracle, that is the estimator which uses the precise knowledge of the non-vanishing components of β_0. Knight and Fu (2000) showed that for $0 < q < 1$ the bridge estimator has the oracle property using a particular tuning parameter λ_n, while Zou (2006) demonstrated that the Lasso can not have it. This author also showed the oracle property for the adaptive Lasso, which determines the estimator minimizing the objective function

$$L(\beta) = \sum_{i=1}^{n} \bigl(Y_i - x_i^T \beta\bigr)^2 + \lambda_n \sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j|^{\gamma}}. \tag{9.3}$$

Here $\beta = (\beta_1, \dots, \beta_p)^T$ and $\tilde{\beta} = (\tilde{\beta}_1, \dots, \tilde{\beta}_p)^T$ denotes a preliminary estimate of β_0 which satisfies certain regularity assumptions. Fan and Peng (2004), Kim et al. (2008), Huang et al. (2008a) and Huang et al. (2008b) derived generalizations of the aforementioned results in the case where the number of parameters is increasing with the sample size, in particular for the case where $p > n$, which will not be considered in this paper.

9.3 Penalizing Estimation Under Heteroscedasticity

In order to investigate the behaviour of the commonly used penalized estimators under heteroscedasticity, we consider the following (heteroscedastic) linear regression model

$$Y = X\beta_0 + \Sigma(\beta_0)\varepsilon, \tag{9.4}$$

where $Y = (Y_1, \dots, Y_n)^T$ is an n-dimensional vector of observed random variables, X is a $n \times p$-matrix of covariates, β_0 is a p-dimensional vector of unknown parameters,

$$\Sigma(\beta_0) = \operatorname{diag}\bigl(\sigma(x_1, \beta_0), \dots, \sigma(x_n, \beta_0)\bigr)$$

is a positive definite matrix and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ is a vector of independent random variables with $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = 1$ for $i = 1, \dots, n$. Note that $\sigma^2(x_i, \beta_0)$ is the variance of the observation Y_i $(i = 1, \dots, n)$. We assume that the model is sparse in the sense that the parameter β_0 can be decomposed as $\beta_0 = (\beta_0(1)^T, \beta_0(2)^T)^T$, where $\beta_0(1) \in \mathbb{R}^k$ and $\beta_0(2) = 0 \in \mathbb{R}^{p-k}$, but it is not known which components of the vector β_0 vanish (this knowledge is called the oracle). Without loss of generality, it is assumed that the k nonzero components are given by the vector $\beta_0(1)^T$. In the following discussion, $x_1^T, \dots, x_n^T$ denote the rows of the matrix X. Moreover, the matrix of covariates is partitioned according to β_0, that is

$$X = \big(X(1), X(2)\big),$$

where $X(1) \in \mathbb{R}^{n \times k}$ and $X(2) \in \mathbb{R}^{n \times (p-k)}$. The rows of the matrix $X(j)$ are denoted by $x_1(j)^T, \dots, x_n(j)^T$ for $j = 1, 2$. We assume that the matrix X is not random but note that for random covariates all results presented in this contribution hold conditionally on the predictor X.

We will also investigate estimators, which take information regarding the scale function into account. For the sake of brevity, we restrict ourselves to estimators of the form

$$\hat{\beta}_{\mathrm{lse}} = \operatorname*{argmin}_{\beta} \left[\sum_{i=1}^{n} \big(Y_i - x_i^T \beta\big)^2 + \lambda_n P(\beta, \tilde{\beta}) \right],$$

$$\hat{\beta}_{\mathrm{wlse}} = \operatorname*{argmin}_{\beta} \left[\sum_{i=1}^{n} \left(\frac{Y_i - x_i^T \beta}{\sigma(x_i, \bar{\beta})} \right)^2 + \lambda_n P(\beta, \tilde{\beta}) \right] \tag{9.5}$$

in the linear heteroscedastic model (9.4). Here $\bar{\beta}$ and $\tilde{\beta}$ denote preliminary estimates of the parameter β_0 and $P(\beta, \tilde{\beta})$ is a penalty function. The estimators $\hat{\beta}_{\mathrm{lse}} = (\hat{\beta}_{\mathrm{lse}}^T(1), \hat{\beta}_{\mathrm{lse}}^T(2))^T$ and $\hat{\beta}_{\mathrm{wlse}} = (\hat{\beta}_{\mathrm{wlse}}^T(1), \hat{\beta}_{\mathrm{wlse}}^T(2))^T$ are decomposed in the same way as the parameter β_0, where the first part corresponds to the k non vanishing coefficients of β_0. We are particularly interested in the cases

$$P(\beta, \tilde{\beta}) = P(\beta) = \|\beta\|_q^q \quad (0 < q \leq 1),$$

$$P(\beta, \tilde{\beta}) = \sum_{j=1}^{p} |\beta_j| |\tilde{\beta}_j|^{-\gamma} \quad (\gamma > 0),$$

corresponding to bridge regression (with the special case of Lasso for $q = 1$) and the adaptive Lasso, respectively, where $\|\beta\|_q = (\sum_{j=1}^{p} |\beta_j|^q)^{1/q}$. The subscripts ‘lse’ and ‘wlse’ correspond to ‘ordinary’ and ‘weighted’ least squares regression, respectively. We mention again that for bridge regression with $q < 1$ the objective functions are not convex in β and there may exist multiple minimizing values. In that case, the argmin is understood as an arbitrary minimizing value, and all results stated here are valid for any such value.

9.3.1 Ordinary Penalized Least Squares Estimators

For regularizing parameters satisfying $\lambda_n/\sqrt{n} \to \lambda_0 > 0$, it was shown in Wagener and Dette (2012) that under heteroscedasticity the classical Lasso performs *conservative* model selection in the same way as in the homoscedastic case, that is

$$\lim_{n\to\infty} P\big(\{1,\ldots,k\} \subset \{j \mid \hat{\beta}_j \neq 0\}\big) = 1 \tag{9.6}$$

and

$$\lim_{n\to\infty} P(\hat{\beta}_j = 0) = c_j \quad \text{for all } j > k \tag{9.7}$$

for some constants $c_1, \ldots, c_{p-k} \in (0,1)$, where $\hat{\beta}_j$ denotes the jth component of the Lasso estimator $\hat{\beta}_{\text{lse}}$ (see, e.g., Leeb and Pötscher 2005). This means that the estimator $\hat{\beta}_{\text{lse}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ estimates a vanishing component of β_0 as 0 with positive probability. On the other hand, conditional on the event that β_{lse} is of the form $(\delta^T, 0^T)$ for some vector $\delta \in \mathbb{R}^k$, the asymptotic covariance matrix of the Lasso estimator

$$\sqrt{n}\big(\hat{\beta}_{\text{lse}}(1) - \beta_0(1)\big)$$

of the nonzero parameters is given by $C_{11}^{-1} B_{11} C_{11}^{-1}$, where the matrices B_{11} and C_{11} are the upper $k \times k$ blocks of the limits

$$\frac{1}{n} X^T X \to C = \begin{pmatrix} C_{11} & C_{21}^T \\ C_{21} & C_{22} \end{pmatrix} > 0, \tag{9.8}$$

and

$$\frac{1}{n} X^T \Sigma(\beta_0)^2 X \to B = \begin{pmatrix} B_{11} & B_{21}^T \\ B_{21} & B_{22} \end{pmatrix} > 0,$$

respectively. Note that results of this type require the existence of the limits and several other technical assumptions. We refer to Wagener and Dette (2012) for a precise formulation of the necessary assumptions such that these statements are true. The asymptotic covariance matrix of the standardized statistic $\sqrt{n}(\hat{\beta}_{\text{lse}}(1) - \beta_0(1))$ is the same as for the ordinary least squares (OLS) estimator in heteroscedastic linear models. This covariance is not the best one achievable, as under heteroscedasticity the OLS estimator is dominated by a generalized LS estimator. Additionally, the estimator is biased if $\lambda_0 \neq 0$.

Similarly, if $\lambda_n/n^{q/2} \to \lambda_0 > 0$, the bridge estimator also performs conservative model selection. Again the asymptotic covariance matrix of the statistic $\sqrt{n}(\hat{\beta}_{\text{lse}}(1) - \beta_0(1))$ is given by $C_{11}^{-1} B_{11} C_{11}^{-1}$ and is suboptimal. However, the estimator is unbiased in contrast to the Lasso estimator.

Finally, if $p < n$, the canonical choice of the preliminary estimate $\tilde{\beta}$ for the adaptive Lasso is the ordinary least squares estimator. In this case, if $\lambda_n n^{(\gamma-1)/2} \to \lambda_0 > 0$, the adaptive Lasso performs conservative model selection, is asymptotically unbiased, but again the asymptotic covariance matrix is not the optimal one.

If the constant c_j in (9.7) is exactly 1, that is

$$\lim_{n\to\infty} P(\hat{\beta}_j = 0) = 1 \quad \text{for all } j > k \tag{9.9}$$

and

$$\lim_{n\to\infty} P(\hat{\beta}_j \neq 0) = 1 \quad \text{for all } j \leq k. \tag{9.10}$$

the estimator $\hat{\beta}$ of the parameter β_0 in model (9.4) is called *consistent* for model selection. This means that the estimator $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ estimates a vanishing component of β_0 as 0 with probability 1. Whether an estimator performs consistent or conservative model selection, respectively, depends on the choice of the tuning parameter λ_n. A 'larger' value of λ_n usually yields consistent model selection while a 'smaller' value yields conservative model selection. It turns out that for the Lasso the goal of consistent model selection contradicts to the property of estimating the non-zero coefficients of the parameter β at the optimal rate $1/\sqrt{n}$. For example, if $\lambda_n/n \to 0$ and $\lambda_n/\sqrt{n} \to \infty$, it can be shown that the estimator $\hat{\beta}_{\mathrm{lse}}$ satisfies

$$\frac{n}{\lambda_n}(\hat{\beta}_{\mathrm{lse}} - \beta_0) \xrightarrow{P} \operatorname{argmin}(V),$$

where the function V is given by

$$V(u) = u^T C u + \sum_{j=1}^{k} u_j \operatorname{sgn}(\beta_{0,j}) + \sum_{j=k+1}^{p} |u_j|,$$

where $\beta_0(1) = (\beta_{0,1}, \ldots, \beta_{0,k})^T$ and the matrix C is defined in (9.8). Moreover, the Lasso performs consistent model selection (even under heteroscedasticity) if the strong *irrepresentable* condition

$$\left|C_{21}C_{11}^{-1} \operatorname{sgn}\bigl(\beta_0(1)\bigr)\right| < \mathbb{1}_{p-k} \tag{9.11}$$

(compare, e.g., Zhao and Yu 2006) is satisfied. Here the inequality and the absolute value are understood component-wise and $\mathbb{1}_{p-k}$ denotes a $p-k$ dimensional vector with all entries given by 1. However, the estimator of the non vanishing parameters does not converge with the optimal rate $1/\sqrt{n}$.

On the other hand, the bridge and adaptive estimator perform consistent model selection and are simultaneously able to estimate the non-vanishing components of the vector β with the optimal rate, if the regularizing parameters satisfy $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n/n^{q/2} \to \infty$ (for the bridge estimator) and $\lambda_n/\sqrt{n} \to 0$, $\lambda_n/n^{(1-\gamma)/2} \to \infty$ (for the adaptive Lasso with a preliminary estimator satisfying $\sqrt{n}(\tilde{\beta} - \beta_0) = O_p(1)$). Both estimators are unbiased but have a non-optimal variance, that is

$$\sqrt{n}\bigl(\hat{\beta}_{\mathrm{lse}}(1) - \beta_0(1)\bigr) \to \mathcal{N}\bigl(0, C_{11}^{-1} B_{11} C_{11}^{-1}\bigr).$$

Due to the lack of convexity of the optimization criterion it is extremely difficult to calculate the bridge estimator in the case $0 < q < 1$. However, the optimization problem corresponding to the adaptive Lasso is convex (but requires a preliminary estimator).

9.3.2 Weighted Penalized Least Squares Estimators

In order to construct an oracle estimator, Wagener and Dette (2012) proposed to use a preliminary estimator, say $\bar{\beta}$, to estimate $\sigma(x_i^T, \beta_0)$ and apply a weighted penalized least squares regression to identify the non-zero components of β_0. This estimator has to satisfy a mild consistency property. The corresponding objective function is defined in (9.5) and the asymptotic properties of the resulting estimator of the nonzero components are described in the following paragraph. For the sake of brevity and transparency, we only mention the mathematical assumptions which are necessary to understand the differences between the different procedures. For detailed discussion of the asymptotic theory, we refer to the work of Wagener and Dette (2012). Roughly speaking the properties of the estimators can be summarized as follows. The weighted bridge and the weighted adaptive Lasso estimator can be tuned to perform consistent model selection and estimation of the non zero parameters with the optimal rate simultaneously. Moreover, the corresponding standardized estimator of the non-vanishing components is asymptotically unbiased and normal distributed with optimal covariance matrix. Different assumptions regarding the tuning parameters and the initial estimators are necessary for statements of this type and some details are given in the following discussion.

Weighted Lasso If

$$\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$$

and the preliminary estimator $\bar{\beta}$ (required for estimating the variance structure) satisfies

$$b_n(\bar{\beta} - \beta_0) = O_p(1) \tag{9.12}$$

for some sequence $0 < b_n \to \infty$ such that $b_n/n^{1/4} \to \infty$, the weighted Lasso estimator $\hat{\beta}_{\text{wlse}}$ performs conservative model selection whenever $\lambda_0 \neq 0$. The standardized estimator $\sqrt{n}(\hat{\beta}_{\text{wlse}}(1) - \beta_0(1))$ of the non-zero components is asymptotically normal distributed with expectation $D_{11}^{-1}\lambda_0 \operatorname{sgn}(\beta_0(1))/2$ and covariance matrix D_{11}^{-1}, where the matrix $D_{11} \in \mathbb{R}^{k\times k}$ is the upper block in the limit matrix

$$\frac{1}{n}X^T \Sigma(\beta_0)^{-2} X \to D = \begin{pmatrix} D_{11} & D_{21}^T \\ D_{21} & D_{22} \end{pmatrix} > 0.$$

This means that the estimator $\hat{\beta}_{\text{wlse}}(1)$ has the optimal asymptotic variance but its standardized version is asymptotically biased.

For

$$\lambda_n/n \to 0, \quad \lambda_n/\sqrt{n} \to \infty,$$

the weighted Lasso estimator performs consistent model selection if the condition

$$\left|D_{21} D_{11}^{-1} \operatorname{sgn}\bigl(\beta_0(1)\bigr)\right| < \mathbb{1}_{p-k}$$

is satisfied. Note that this condition is an analogue of the strong irrepresentable condition (9.11). In this case, the estimator of the nonvanishing components does not converge at the optimal rate.

Weighted Adaptive Lasso Note that the weighted adaptive Lasso requires two preliminary estimators which can but must not necessarily coincide. One estimator $\tilde{\beta}$ is used in the penalty and a further estimator $\bar{\beta}$ is required for the parameters in the variance function $\sigma^2(x, \beta)$.

If the preliminary estimators $\bar{\beta}$ and $\tilde{\beta}$ of β_0 satisfy (9.12) and

$$a_n(\tilde{\beta} - \beta_0) \overset{\mathcal{D}}{\to} Z \tag{9.13}$$

for some positive sequence $a_n \to \infty$, such that $P(Z = 0) = 0$ and

$$\lambda_n a_n^{\gamma} / \sqrt{n} \to \lambda_0 \geq 0,$$

the weighted adaptive Lasso estimator $\hat{\beta}_{\text{wlse}}$ performs conservative model selection and the standardized estimator $\sqrt{n}(\hat{\beta}_{\text{wlse}}(1) - \beta_0(1))$ of the nonzero components is asymptotically normal distributed with mean 0 and covariance matrix D_{11}^{-1}.

If

$$\lambda_n/\sqrt{n} \to 0, \quad \lambda_n/\sqrt{n} a_n^{\gamma} \to \infty$$

for some sequence $a_n \to \infty$ and the preliminary estimators $\bar{\beta}$ and $\tilde{\beta}$ of β_0 satisfy (9.12) and $a_n(\tilde{\beta} - \beta_0) = O_p(1)$, respectively, the weighted adaptive Lasso estimator $\hat{\beta}_{\text{wlse}}$ performs consistent model selection, is asymptotically unbiased and converges weakly with the optimal rate and covariance matrix, that is

$$\sqrt{n}\bigl(\hat{\beta}_{\text{wlse}}(1) - \beta_0(1)\bigr) \to \mathcal{N}\bigl(0, D_{11}^{-1}\bigr).$$

Weighted Bridge Estimation We assume that $q \in (0, 1)$. If

$$\lambda_n / n^{q/2} \to \lambda_0 \geq 0$$

and the preliminary estimator $\bar{\beta}$ of β_0 satisfies (9.12), the weighted bridge estimator $\hat{\beta}_{\text{wlse}}$ performs conservative model selection and the standardized estimator $\sqrt{n}(\hat{\beta}_{\text{wlse}}(1) - \beta_0(1))$ of the nonzero components is asymptotically normal distributed with mean 0 and covariance matrix D_{11}^{-1}.

If

$$\lambda_n/\sqrt{n} \to 0, \quad \lambda_n/n^{q/2} \to \infty,$$

the weighted bridge estimator performs consistent model selection and estimates the nonzero parameters with the optimal rate and optimal covariance matrix D_{11}^{-1}.

9.4 Finite Sample Properties

In this section, we present a small simulation study and illustrate the differences between the various estimators in a data example. We mention again that despite of their appealing theoretical properties bridge estimators are extremely difficult to calculate and therefore we concentrate on a comparison of the Lasso and adaptive Lasso. These estimators can be calculated by convex optimization, and we used the package "penalized" available for R on http://www.R-project.org (R Development Core Team 2008) to perform all computations. The numerical results presented here are taken from Wagener and Dette (2012). Further finite sample results can be found in this reference.

In all examples, the data were generated using a linear model (9.4). The errors ε were i.i.d. standard normal and the matrix Σ was a diagonal matrix with entries $\sigma(x_i, \beta_0)$ on the diagonal, where σ was given by one of the following functions:

$$\begin{aligned}
&\text{(a)} \quad \sigma(x_i, \beta_0) = \frac{1}{2}\sqrt{x_i^T \beta_0},\\
&\text{(b)} \quad \sigma(x_i, \beta_0) = \frac{1}{4}|x_i^T \beta_0|,\\
&\text{(c)} \quad \sigma(x_i, \beta_0) = \frac{1}{20}\exp|x_i^T \beta_0|,\\
&\text{(d)} \quad \sigma(x_i, \beta_0) = \frac{1}{50}\exp\left(x_i^T \beta_0\right)^2.
\end{aligned}$$

The different factors were chosen in order to generate data with comparable variance in each of the four models. The tuning parameter λ_n was chosen by fivefold generalized cross validation performed on a training data set. For the preliminary estimator $\tilde{\beta}$ we used the OLS estimator and for $\bar{\beta}$ the unweighted Lasso estimator. All reported results are based on 100 simulation runs. The design matrix was generated having independent normally distributed rows and the covariance between the i-th and j-th entry in each row was $0.5^{|i-j|}$. The sample size was chosen by $n = 60$. We consider the model (9.4) with parameter $\beta = (3, 1.5, 2, 0, 0, 0, 0, 0)$ (see Zou 2006). The model selection performance of the estimators is presented in Table 9.1, where we show the mean of the correctly zero and correctly non-zero estimated parameters. In the ideal case these should be 5 and 3, respectively. It can be seen from Table 9.1 that the adaptive Lasso always performs better with respect to model selection than the Lasso. This confirms the asymptotic properties described in the previous section. The model selection performance of the weighted and unweighted adaptive Lasso are comparable. In Table 9.2, we present the mean squared error of the estimates for the non-vanishing components $\beta_1, \beta_2, \beta_3$. In terms of this criterion, the weighted versions of the estimators nearly always do a (in some cases substantially) better job than their unweighted counterparts.

We conclude this chapter by illustrating the differences between the two estimators $\hat{\beta}_{\text{lse}}$ and $\hat{\beta}_{\text{wlse}}$ with reanalyzing the diabetes data considered in Efron et al. (2004). The data consist of a response variable Y which is a quantitative measure of

Table 9.1 Mean number of correctly zero and correctly nonzero estimated parameters in model (9.4) with $\beta = (3, 1.5, 2, 0, 0, 0, 0, 0)$

		σ			
		(a)	(b)	(c)	(d)
Lasso	$=0$	1.67	3.33	1.58	2.64
	$\neq 0$	3	3	2.99	3
Adaptive Lasso	$=0$	4.51	4.32	2.95	4.48
	$\neq 0$	3	3	2.95	3
Weighted Lasso	$=0$	0.97	1.53	0.67	0.43
	$\neq 0$	3	3	3	3
Weighted adaptive Lasso	$=0$	3.97	4.09	3.29	3.91
	$\neq 0$	3	3	3	3

Table 9.2 Mean squared error of the estimators of the nonzero coefficients in model (9.4) with $\beta = (3, 1.5, 2, 0, 0, 0, 0, 0)$

		σ			
		(a)	(b)	(c)	(d)
Lasso	β_1	0.0308	0.0682	0.3480	0.0692
	β_2	0.0306	0.0374	0.2461	0.0784
	β_3	0.0322	0.0484	0.3483	0.1141
Adaptive Lasso	β_1	0.0293	0.0593	0.3514	0.0668
	β_2	0.0330	0.0393	0.3241	0.1027
	β_3	0.0285	0.0416	0.3871	0.1126
Weighted Lasso	β_1	0.0215	0.0424	0.1431	0.2004
	β_2	0.0171	0.0133	0.0458	0.0174
	β_3	0.0191	0.0202	0.1086	0.0780
Weighted adaptive Lasso	β_1	0.0193	0.0152	0.0944	0.1953
	β_2	0.0168	0.0069	0.0293	0.0134
	β_3	0.0165	0.0080	0.0864	0.0763

diabetes progression one year after baseline and of ten covariates (age, sex, body mass index, average blood pressure and six blood serum variables). It includes $n = 442$ observations.

First, we calculated the unweighted Lasso estimate $\hat{\beta}_{\text{lse}}$ using a cross-validated (conservative) tuning parameter λ_n. This solution excluded three covariates from the model (age and two of the six blood serum variables). In a next step we calculated the resulting residuals

$$\varepsilon = Y - X\hat{\beta}_{\text{lse}},$$

which are plotted in the upper left panel of Fig. 9.1.

This picture suggests a heteroscedastic nature of the residuals. In fact the hypothesis of homoscedastic residuals was rejected by the test of Dette and Munk (1998) which had a p-value of 0.006. Next, we computed a local linear fit of the squared

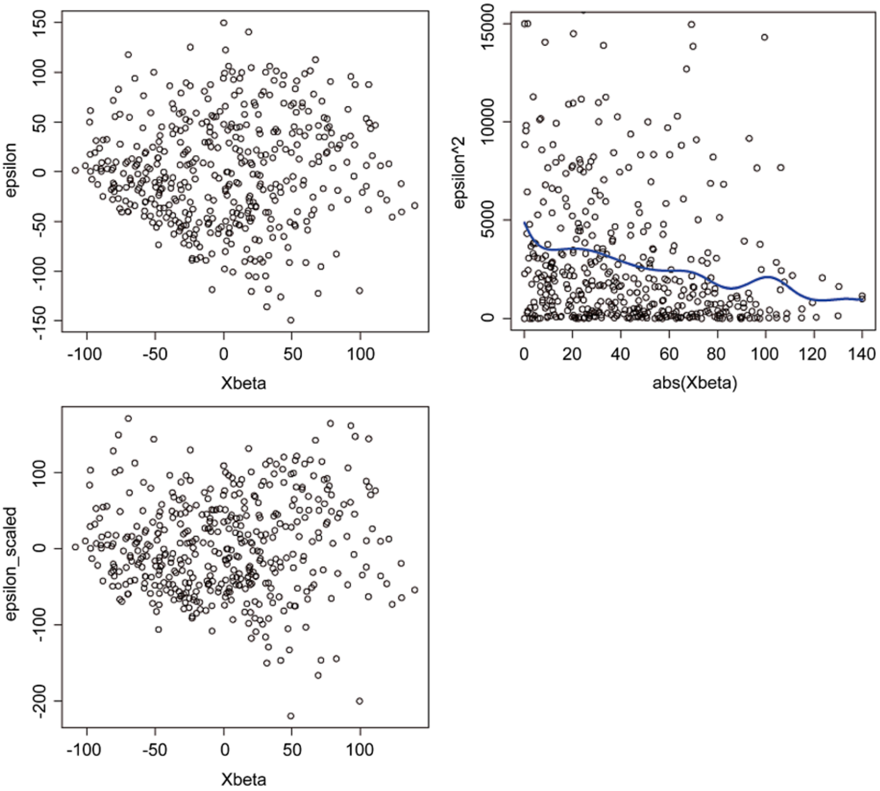


Fig. 9.1 *Upper left*: Lasso residuals, *upper right*: Squared residuals together with local polynomial estimate, *lower left*: rescaled residuals

residuals in order to estimate the conditional variance $\sigma(x_i^T \beta)$ of the residuals. The upper right panel of Fig. 9.1 presents the squared residuals plotted against its absolute values $|x_i^T \hat{\beta}_{\text{lse}}|$ together with the local linear smoother, say $\hat{\sigma}^2$. In the lower left panel of Fig. 9.1 we present the rescaled residuals $\tilde{\varepsilon}_i = (Y_i - x_i^T \hat{\beta}_{\text{lse}})/\hat{\sigma}(|x_i^T \hat{\beta}_{\text{lse}}|)$. These look "more homoscedastic" than the unscaled residuals and the test of Dette and Munk (1998) has a p-value of 0.514, thus not rejecting the hypothesis of homoscedasticity. The weighted Lasso estimator $\hat{\beta}_{\text{wlse}}$ was calculated by (9.5) on the basis of the "nonparametric" weights $\hat{\sigma}(x_i^T \hat{\beta}_{\text{lse}})$ and the results are depicted in Table 9.3. In contrast to $\hat{\beta}_{\text{lse}}$, the weighted Lasso only excludes one variable from the model, namely the blood serum HDL if λ_n is chosen by cross-validation.

9.5 Conclusions

In this paper, we presented several penalized least squares methods for estimation in sparse high-dimensional linear regression models. We explain the terminology

Table 9.3 The Lasso estimators $\hat{\beta}_{\text{lse}}$ and $\hat{\beta}_{\text{wlse}}$ for the tuning parameter λ_n selected by cross-validation

	Intercept	Age	Sex	BMI	BP	TC	LDL	HDL	TCH	LTG	GLU
$\hat{\beta}_{\text{lse}}$	152.1	0.0	−186.5	520.9	291.3	−90.3	0.0	−220.2	0.0	506.6	49.2
$\hat{\beta}_{\text{wlse}}$	183.8	−110.3	−271.3	673.3	408.3	84.1	−547.6	0.0	449.4	213.7	138.5

of conservative and consistent model selection. It is demonstrated that under heteroscedasticity these estimators can be modified, such that they are as efficient as oracle estimators. On the other hand, classical methods like Lasso or adaptive Lasso yield consistent estimators but with sub-optimal rates.

Acknowledgements This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt C1) of the German Research Foundation (DFG). The authors would also like to thank Torsten Hothorn for pointing out some important references on the subject.

References

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, *37*, 373–384.

Candes, E., & Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, *35*, 2313–2351.

Dette, H., & Munk, A. (1998). Testing heteroscedasticity in nonparametric regression. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *60*, 693–708.

Efron, B., Hastie, T., & Tibshirani, R. (2004). Least angle regression (with discussion). *The Annals of Statistics*, *32*, 407–451.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Fan, J., & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, *32*, 928–961.

Frank, I. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, *35*, 109–148.

Huang, J., Horowitz, J. L., & Ma, S. (2008a). Asymptotic properties of bridge estimators in sparse high dimensional regression models. *The Annals of Statistics*, *36*, 587–613.

Huang, J., Ma, S., & Zhang, C. (2008b). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, *18*, 1603–1618.

Kim, Y., Choi, H., & Oh, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, *103*, 1665–1673.

Knight, F., & Fu, W. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics*, *28*, 1356–1378.

Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory*, *21*, 21–59.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*, 267–288.

Wagener, J., & Dette, H. (2012). Bridge estimators and the adaptive Lasso under heteroscedasticity. *Mathematical Methods of Statistics*. doi:10.3103/S1066530712020032

Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *67*, 301–320.

Chapter 10
Bayesian Smoothing, Shrinkage and Variable Selection in Hazard Regression

Susanne Konrath, Ludwig Fahrmeir, and Thomas Kneib

10.1 Introduction

In modern regression analysis, penalized likelihood approaches and related Bayesian concepts have emerged as general tools both for semiparametric smoothing of nonlinear covariate effects and for the regularization of high-dimensional vectors of linear covariate effects. Approaches for combining regularization and smoothing, as increasingly required in applications are quite sparse and have not yet been developed for hazard regression.

In this contribution, we extend the Bayesian regularization approaches described in Kneib et al. (2010) and Fahrmeir et al. (2010) for Gaussian and exponential family regression to Cox-type hazard regression models with predictors including high-dimensional linear covariate effects, time-varying and nonlinear effects of covariates and—for a full likelihood approach—the baseline hazard rate.

Regularization of linear predictors relies on hierarchical prior formulations, where the conditional prior for the regression coefficients is Gaussian with a suitable (mixture) hyperprior on the variance. In particular, all marginal priors for regression coefficients can be expressed as scale mixtures of normals. With the exception of the SCAD penalty (Fan and Li 2002), these priors comprise all Bayesian versions of penalties suggested for high-dimensional Cox models: The ridge penalty (van Houwelingen et al. 2006), the lasso (Tibshirani 1997), the adaptive lasso (Zou

S. Konrath (✉) · L. Fahrmeir
Department of Statistics, Ludwig-Maximilians-University Munich, Ludwigstraße 33,
80539 Munich, Germany
e-mail: susanne.konrath@stat.uni-muenchen.de

L. Fahrmeir
e-mail: ludwig.fahrmeir@stat.uni-muenchen.de

T. Kneib
Georg-August-University Göttingen, Platz der Göttinger Sieben 5, 37073 Göttingen, Germany
e-mail: tkneib@uni-goettingen.de

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_10, © Springer-Verlag Berlin Heidelberg 2013

2006) and the elastic net, initially proposed for Gaussian models by Park and Hastie (2007) and implemented for high-dimensional Cox models by Benner et al. (2010). Their article also provides a good comparative review of these penalized likelihood approaches. Bridge and (adaptive) lasso penalties are also reviewed and extended to high-dimensional heteroscedastic linear models in the contribution by Dette and Wagener, Chap. 9. In addition, a spike and slab hyperprior for variances of Gaussian regularization priors, suggested by Ishwaran and Rao (2003, 2005) for Gaussian linear models, is also covered by our general formulation. Smoothness priors for semiparametric components such as B-splines for the (log-) baseline hazard rate, nonlinear and time varying effects of covariates are formulated through conditionally Gaussian densities for basis coefficients, given variance or smoothing parameters, as in Hennerfeind et al. (2006) for hazard regression and Brezger and Lang (2006) for exponential family regression.

Thus, conditionally Gaussian regularization and smoothness priors for high-dimensional regression coefficients and basis function coefficients are the key to a unified joint modelling framework. Marginal priors for these coefficients differ through further priors for variances in the hierarchical formulation, inducing the desired smoothness or sparseness properties.

Based on the scale mixture representation, we develop unified MCMC algorithms for inference in hazard regression. In particular, samples from non-Gaussian full conditionals for high-dimensional regression coefficients and for basis function coefficients can be drawn in a unified computationally efficient way based on Metropolis Hastings (MH) steps with iteratively weighted least squares (IWLS) proposals requiring no further tuning. As an advantage of MCMC based inference, standard errors for coefficients with shrinkage priors are easily available and estimation of shrinkage parameters is integrated within the sampling scheme. Moreover, extensions to adaptive priors are possible without great effort. Variable selection can also be included into the inferential process: For the ridge and lasso prior, we suggest an empirical thresholding procedure, while MCMC with spike and slab prior for shrinkage variances allows directly estimating inclusion probabilities, see Sect. 10.3.4.

The presented methods based on the full likelihood are implemented in the free software BayesX (http://www.bayesx.org), for partial likelihood inference the used R-functions are available on request from the first author.

The majority of previous Bayesian approaches to regularize high-dimensional linear predictors have focused on Gaussian or, to a smaller extent, on exponential family regression models. The Bayesian lasso prior (Park and Casella 2008), the elastic net prior (Li and Lin 2010), the spike and slab priors of George and McCulloch (1993), Smith and Kohn (1996), and Ishwaran and Rao (2005), as well as others discussed in Griffin and Brown (2005) have all been proposed and developed for high-dimensional regression models with Gaussian responses. In fact, efficient posterior inference for some of them crucially depends on the Gaussian assumption and is not easily extendable to non-Gaussian regression models.

Joint regularization of high-dimensional linear effects and smoothness prior approaches are still rare. Panagiotelis and Smith (2008) focus on function selection in Gaussian additive models, but high-dimensional linear effects could easily be incorporated. Bayesian shrinkage priors in combination with smoothing are investigated for Gaussian models in Kneib et al. (2010). A recent spike and slab prior for function selection in structured additive regression models has been proposed in Scheipl (2011) and Scheipl et al. (2012). An extension to hazard regression would be desirable.

The rest of this chapter is organized as follows: Sects. 10.2 and 10.3 introduce different types of hazard regression models and Bayesian regularization priors. In particular, scale mixture representations of the ridge, the lasso, and the spike and slab prior will be introduced along with regularization priors for penalized spline smoothing. Section 10.4 discusses posterior inference based on MCMC simulations. Sections 10.5 and 10.6 are devoted to simulations and applications to demonstrate the flexibility and applicability of the proposed methodology. Finally, Sect. 10.7 contains a conclusion with directions of future research. Additional results are provided in the electronic supplement available from http://www.uni-goettingen.de/de/304963.html.

10.2 Survival Models and Likelihoods

Survival models are regression models for analysing the influence of covariates on survival times or, more generally, on times until a certain event of interest occurs. Most popular are continuous-time hazard rate regression models: The influence of covariates $\boldsymbol{x}$ on survival time $T \geq 0$ is specified through a regression model for the hazard rate

$$\lambda\big(t \mid \mathbf{x}\big) = \lim_{h \to 0} \frac{1}{h} P\big(t \leq T < t + h \mid T \geq t, \mathbf{x}\big).$$

The hazard rate can be interpreted as the instantaneous rate of an event in the interval $[t, t+h)$ given survival up to t. A typical feature of survival times T_i for individuals $i = 1, \ldots, n$ is that a percentage of them is not completely observed but censored or truncated. The most common type of incomplete survival data are right censored observations as in our application in Sect. 10.6. This censoring schemes can be formalised by censoring times C_i. Instead of observing true survival times T_i, only $t_i = \min(T_i, C_i)$ is observed together with the censoring indicator $d_i = I\ (T_i \leq C_i)$, i.e. $d_i = 1$ if $T_i \leq C_i$ and $d_i = 0$ else. A common assumption is that survival times and censoring times are independent, and that censoring is noninformative, i.e., the distribution of C_i does not depend on parameters of interest contained in the distribution of T_i. For the shrinkage and smoothness concepts considered in the next sections, we additionally assume that continuous covariates are standardized in advance, thus avoiding adjustment of shrinkage priors for different covariate scales.

In Cox's proportional hazard model (Cox 1972), the hazard rate for individual i is given as

$$\lambda_i(t) = \lambda_0(t) \exp\big(\mathbf{x}_i' \boldsymbol{\beta}\big) = \exp\big(\log \lambda_0(t) + \mathbf{x}_i' \boldsymbol{\beta}\big)$$

for $t > 0$ with predictor $\eta_i(t) = \log \lambda_0(t) + \mathbf{x}_i'\boldsymbol{\beta}$, where $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$ is a vector of time independent covariates. The baseline hazard $\lambda_0(t)$ is left unspecified. This model separates time-dependence and cocariate effects, making the model easy to interpret but also too resrtictive in some applications. Restrictions can be relaxed by more general predictors as mentioned below, including time-varying effects or covariates. Under independent and noninformative right-censoring the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, $p < n$, are usually estimated via maximization of the partial likelihood (Cox 1972)

$$pL(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sum_{k \in R(t_i)} \exp(\mathbf{x}_k'\boldsymbol{\beta})} \right)^{d_i}, \tag{10.1}$$

where $R(t_i)$ denotes the risk set at time t_i as the set of all individuals who are still under study at a time just prior to t_i. The cumulative baseline hazard $\Lambda_0(t) = \int_0^t \lambda_0(u)\,du$ is then estimated in a second step by the Breslow estimator, see, for example, Lin (2007). More details on the Cox model and other regression models for survival data can be found, for example, in Klein and Moeschberger (2003).

To use the full likelihood

$$L(\boldsymbol{\beta}, \lambda) = \prod_{i=1}^{n} \lambda_0(t_i) \exp\big(\mathbf{x}_i'\boldsymbol{\beta}\big)^{d_i} \exp\left(- \int_0^{t_i} \lambda_0(s) \exp\big(\mathbf{x}_i'\boldsymbol{\beta}\big)\, ds \right), \tag{10.2}$$

requires, in contrast to the partial likelihood, the specification of the baseline hazard function $\lambda_0(t)$ because the log-baseline hazard is explicitly included in the likelihood. Based on the full likelihood joint inference for covariate effects and the baseline hazard becomes feasible and obtaining a full probabilistic framework is useful if modeling of individual hazards and predictions are of interest. Besides simple parametric models also more flexible specifications for the shape of the baseline hazard are feasible. For example, we model the logarithm $g_0(t) = \log \lambda_0(t)$ of the baseline hazard through a polynomial spline with smoothness prior for the basis coefficients (details in Sect. 10.3.5).

In addition the predictor $\eta_i(t)$ can further be extended to include different kinds of nonlinear covariate effects, e.g.,

$$\lambda_i(t) = \exp\big(\eta_i(t)\big) = \exp\left(g_0(t) + \mathbf{x}_i'\boldsymbol{\beta} + \sum_{j=1}^{q} f_j(z_{ij}) \right), \tag{10.3}$$

where $f_j(\mathbf{z}_j)$ are nonlinear effects of continuous covariates $\mathbf{z}_j$. Further components, such as covariates with linear effects that should not be regularized by a special prior, time-varying effects, nonlinear interactions between two continuous covariates, spatial effects and group- or individual specific effects (frailties) may also be included. Cox-type models with such general forms of structured additive predictors have been suggested in Hennerfeind et al. (2006), however without considering special kinds of regularization of $\boldsymbol{\beta}$, e.g., to deal simultaneously with variable selection. Apart from simple parametric forms of the baseline hazard rate, for example a Weibull model or a piecewise constant function, inserting $\lambda_i(t_i)$ from expression (10.3) into the likelihood (10.2) requires that the integral of the hazard function in

(10.2) has to be evaluated numerically, using, e.g., the trapezoidal rule as in Hennerfeind et al. (2006).

For Bayesian inference, it seems questionable if the partial likelihood (10.1) can be used instead of the genuine full likelihood (10.2) for posterior analysis. Sinha et al. (2003) provide a rigorous justification for the use of the partial likelihood in the Bayesian context, when the (cumulative) baseline hazard is specified through a gamma process prior. We will instead specify $\lambda_0(t)$ through a log-normal process prior. Because gamma and log-normal process priors are closely related from a practical point of view, we argue heuristically that Bayesian inference with extended predictor can again be based on the partial likelihood. Section 10.5 provides empirical evidence for this conjecture.

10.3 Shrinkage and Smoothness Priors

To deal with the problem of variable selection and regularization, including models with more parameters than observations, we consider and compare several shrinkage priors. The priors considered in the following sections can all be represented as scale mixtures of normals, which allow on the one hand treating them in our unifying framework based on conditionally Gaussian structures and is on the other hand very useful for sampling in MCMC inference.

10.3.1 Ridge Prior

The Bayesian version of the ridge penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ is given by the assumption of i.i.d. Gaussian priors for the regression coefficients $\beta_j \mid \lambda \sim_{\text{i.i.d.}} N(0, 1/2\lambda)$, $j = 1, \dots, p$, $\lambda > 0$ with the scale mixture representation

$$\beta_j \mid \tau_j^2 \sim N\big(0; \tau_j^2\big), \quad \tau_j^2 \mid \lambda \sim \delta_{1/2\lambda}\big(\tau_j^2\big).$$

The symbol $\delta_a(t)$ denotes the Kronecker function which is 1 if $t = a$ and 0 if $t \neq a$. For given $\lambda > 0$, posterior mode estimation corresponds to penalized likelihood estimation. Due to conjugacy to the Gaussian family, an additional gamma prior is used for the shrinkage parameter $\lambda \sim \text{Ga}(a_\lambda, b_\lambda)$, $a_\lambda, b_\lambda > 0$, which supports a Gibbs update for this parameter. This enables to determine the shrinkage parameter λ also in cases with very complex predictor in the typical Bayesian fashion as posterior median or mean of the corresponding sample and also entails a more flexible prior for the regression coefficients. The marginalization over λ results in an inverse gamma distribution for the variance parameters $\tau_j^2 \mid a_\lambda, b_\lambda \sim \text{IG}(a_\lambda, b_\lambda/2)$ and further marginalization in a scaled t distribution for the distribution of the regression coefficients

$$\beta_j \mid a_\lambda,\ b_\lambda \sim t\big(\beta_j \mid \text{df} = 2a_\lambda, \text{scale} = \sqrt{b_\lambda/2a_\lambda}\big)$$

given the hyperparameters a_λ, b_λ. Figure 10.1 shows this marginal log-prior in the upper right panel. In the upper left panel the log of the Gaussian marginal prior is displayed which results if there is no prior assumption for λ.

10.3.2 Lasso Prior

Just as well known as ridge regression in the context of collinearity is the lasso regression (Tibshirani 1997) if simultaneous variable selection and estimation should be achieved. The Bayesian version of the lasso $\lambda \sum_{j=1}^{p} |\beta_j|$ can be formulated with i.i.d. centered Laplace priors $\beta_j \mid \lambda \sim_{\text{i.i.d.}} \text{Lap}(0, \lambda)$, $j = 1, \ldots, p$, where $\lambda > 0$ represents the inverse scale parameter of the Laplace distribution with density $p(\beta) \propto \exp(-\lambda|\beta|)$, compare e.g., Park and Casella (2008). Figure 10.1 shows the Laplace log-prior in the univariate case. For full Bayesian inference, it is advantageous to express the Laplace density as a scale mixture of normals introducing a further stage in the hierarchical model formulation:

$$\beta_j \mid \tau_j^2 \sim N\big(0; \tau_j^2\big), \quad \tau_j^2 \mid \lambda^2 \sim_{\text{i.i.d.}} \text{Exp}\left(\frac{\lambda^2}{2}\right).$$

To obtain a data driven penalty, we additionally use again a gamma prior for the squared shrinkage parameter λ^2, i.e., $\lambda^2 \sim \text{Ga}(a_\lambda, b_\lambda)$, $a_\lambda, b_\lambda > 0$ and get the same benefits for the estimation of the shrinkage parameter as mentioned in the Bayesian ridge section. This hierarchy leads to

$$\pi\big(\tau_j^2 \mid a_\lambda, b_\lambda\big) = 0.5 a_\lambda b_\lambda^{-1}\big[0.5\tau_j^2 b_\lambda^{-1} + 1\big]^{-(a_\lambda+1)}$$

as the density of the marginal distributions for the variance parameter. The corresponding marginal density of the regression coefficients (Fig. 10.1, lower left panel) can be expressed as

$$\pi\big(\beta_j \mid a_\lambda, b_\lambda\big) = \frac{1}{\sqrt{\pi}} \frac{a_\lambda 2^{a_\lambda}}{\sqrt{2b_\lambda}} \Gamma\left(a_\lambda + \frac{1}{2}\right) \exp\left(\frac{1}{4}\frac{\beta_j^2}{(2b_\lambda)}\right) D_{-2a_\lambda-1}\left(\sqrt{\frac{\beta_j^2}{2b_\lambda}}\right)$$

with $D_{-2a_\lambda-1}$ as the parabolic cylinder function, see Griffin and Brown (2005, 2010) for details.

10.3.3 Normal Mixture of Inverse Gamma Prior

As a further mixture prior, we consider the normal mixture of inverse gamma (NMIG) distribution suggested for regularizing high-dimensional linear Gaussian regression models by George and McCulloch (1993) and Ishwaran and Rao (2003, 2005). The conditional distribution for the regression coefficients is Gaussian as in the lasso and ridge case, i.e., $\beta_j \mid I_j, \psi_j^2 \sim N(0; \tau_j^2 = I_j \psi_j^2)$, but in contrast the variance parameters τ_j^2 of this distribution are specified through a mixture distribution modeled by the product of the two components, i.e.,

$$I_j \mid \nu_0, \nu_1, \omega \sim (1-\omega)\delta_{\nu_0}(\cdot) + \omega\delta_{\nu_1}(\cdot), \quad \psi_j^2 \mid a_\psi, b_\psi \sim \text{IG}(a_\psi, b_\psi). \qquad (10.4)$$

The first component in (10.4) is an indicator variable with point mass at the values $\nu_0 > 0$ and $\nu_1 > 0$ denoted by the corresponding Kronecker symbols. Therein the parameter ν_0 should have a positive value close to zero and the value of ν_1 is set to 1. The parameter ω controls how likely the binary variable I_j equals ν_1 or ν_0, and therefore it takes on the role of a complexity parameter that controls the size of the models. The assumptions in (10.4) are leading to a continuous bimodal distribution for the variance parameters $\tau_j^2 := I_j \psi_j^2$, given ν_0, ν_1, ω, a_ψ, b_ψ with representation as a mixture of scaled inverse gamma distributions:

$$\pi\left(\tau_j^2 \mid \nu_0, \nu_1, \omega, a_\psi, b_\psi\right) = (1-\omega)\cdot \mathrm{IG}\left(\tau_j^2 \mid a_\psi, \nu_0 b_\psi\right) + \omega \cdot \mathrm{IG}\left(\tau_j^2 \mid a_\psi, \nu_1 b_\psi\right).$$

We assume a beta prior Beta(a_ω, b_ω) for the parameter ω, which reduces to the uniform prior in the special case $a_\omega = b_\omega = 1$ to express an indifferent prior knowledge about the model complexity. With an appropriate choice of the hyperparameters $a_\omega, b_\omega > 0$, it is possible to favor more or less sparse models. Using a continuous prior for ω has several advantages over using a degenerate point mass at zero prior as for example in George and McCulloch (1997) for linear models. First, the update of the variance parameter components can easily be done via Gibbs sampling and no complicated updates are necessary. Furthermore, it is possible to select important model variables via the posterior mean of the corresponding indicators I_j and to simultaneously estimate their values like for λ in the lasso case and the choice of a beta prior for ω allows for a greater amount of adaptiveness in estimating the model size. Besides dealing with the case $p > n$ or multicollinearity, a desirable feature of shrinkage priors used for variable selection is to shrink small effects close to zero, but to shrink significant effects only moderately to prevent them from large bias, see the discussions in Griffin and Brown (2005, 2010) or Zou (2006) for linear models. The NMIG-estimates for relevant covariates should be less biased than in the case of unimodal priors for the regression coefficients because the bimodality supports less penalization of large coefficients.

The marginal density for the variance parameters, after integrating out ω, is the mixture

$$\tau_j^2 \mid \nu_0, \nu_1, a_\psi, b_\psi \sim 0.5 \cdot \mathrm{IG}\left(\tau_j^2 \mid a_\psi, \nu_0 b_\psi\right) + 0.5 \cdot \mathrm{IG}\left(\tau_j^2 \mid a_\psi, \nu_1 b_\psi\right),$$

which corresponds to the conditional density $\pi(\tau_j^2 \mid \nu_0, \nu_1, \omega, a_\psi, b_\psi)$ for the choice $\omega = 0.5$. The prior locations of the two modes are independent of ω and fixed at $\nu_0 b_\psi/(a_\psi + 1)$ and $\nu_1 b_\psi/(a_\psi + 1)$. The marginal distribution for the components of β is a mixture of scaled t-distributions

$$\begin{aligned}\beta_j \mid \nu_0, \nu_1, a_\psi, b_\psi \sim\ & 0.5 \cdot t\left(\beta_j \mid \mathrm{df} = 2a_\psi, \mathrm{scale} = \sqrt{\frac{\nu_0 b_\psi}{a_\psi}}\right) \\ & + 0.5 \cdot t\left(\beta_j \mid \mathrm{df} = 2a_\psi, \mathrm{scale} = \sqrt{\frac{\nu_1 b_\psi}{a_\psi}}\right).\end{aligned}$$

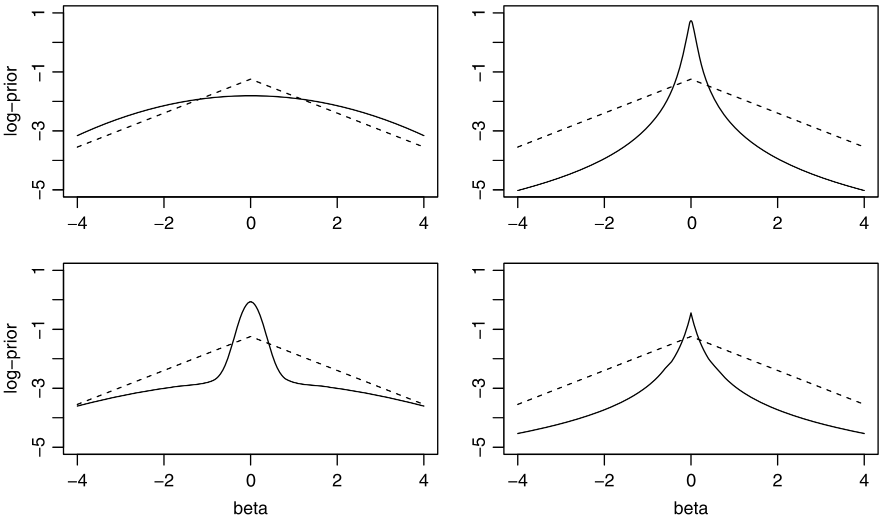


Fig. 10.1 Marginal log-priors (*solid lines*) in comparison to the marginal lasso log-prior (*dashed line*) corresponding to the Ridge penalty (*upper left panel*), the Bayesian ridge penalty (*upper right panel*), the Bayesian lasso penalty (*lower left panel*) and the Bayesian NMIG penalty (*lower right panel*)

The corresponding log-prior is displayed in Fig. 10.1, lower right panel. Additional topics, like derivation of the marginal priors can be found in the electronic supplement.

10.3.4 Variable Selection

Since the Bayesian regularization priors do not share the strong variable selection property of the frequentist lasso, hard shrinkage rules are considered to accomplish variable selection. A first rule is based on the 95 % credible intervals (CI95), obtained from the corresponding sample quantiles of the MCMC samples for the regression coefficient. A second interval criterion is constructed using the sample standard deviation (STD), so that only regression coefficients with zero outside the one standard deviation interval around the posterior mean are included in the final model. On the other hand, as mentioned before in Sect. 10.3.3, the Bayesian NMIG provides a natural criterion to select covariates if the samples of the indicator variables I_j are utilized. Covariates with considerable influence should be assigned to the mixing distribution component corresponding to the indicator with values $\nu_1 = 1$. The more the posterior mean of an indicator variable increases (i. e. the percentage of the values $\nu_1 = 1$ in the sample), the larger is the evidence that the corresponding covariate has non negligible effect. In our simulations and application, we use the intuitive cut value of 0.5 as a selection criterion, i.e., covariates whose corresponding indicator posterior mean exceeds 0.5 are included in the final model, see Sects. 10.5 and 10.6.

10.3.5 Smoothness Priors

In our work, smooth modelling and estimation of nonlinear effects including the (log-) baseline hazard, is based on Bayesian P-splines, compare Lang and Brezger (2004). We illustrate the generic approach using the baseline hazard function as an example. If the partial instead of the full likelihood is used, no assumptions for modeling the baseline hazard are needed but the principle for estimating the remaining smooth components of the predictor stays the same.

The baseline hazard is approximated by a linear combination of $M = s + \ell - 1$ degree ℓ B-spline basis functions $B_1(\cdot), \dots, B_M(\cdot)$ with an appropriate sequence of knots $\xi_1 < \cdots < \xi_s$ from $(t_{\min}, t_{\max})$ with additional boundary knots $0 = \xi_0 < \xi_1$ and $\xi_s < \xi_{s+1} = \infty$. The resulting representation $g_0(t) = \log \lambda_0(t) = B_1(t)\,\gamma_{0,1} + \cdots + B_M(t)\,\gamma_{0,M}$ enables that the baseline hazard function can be expressed as the product of an appropriately defined design matrix $\mathbf{Z}_0$ with rows $\mathbf{z}'_{i0} = (B_1(t_i), \dots, B_M(t_i))$ and a vector $\boldsymbol{\gamma}_0$ of parameters, i.e., $\mathbf{g}_0 = \mathbf{Z}_0\boldsymbol{\gamma}_0$, if $\mathbf{g}_0 = (g_0(t_1), \dots, g_0(t_n))'$ denotes the vector of function evaluations of $g_0(t_i)$.

To guarantee smoothness for the unknown log baseline hazard $g_0(t)$, we assume first or second order random walk smoothness priors for the parameter vector $\boldsymbol{\gamma}_0$

$$\gamma_{0,m} = \gamma_{0,m-1} + u_{0,m} \quad \text{or} \quad \gamma_{0,m} = 2\gamma_{0,m-1} - \gamma_{0,m-2} + u_{0,m},$$

with i.i.d. Gaussian errors $u_{0,m} \sim N(0, \delta_0^2)$ and diffuse priors for the initial values $\pi(\gamma_{0,1}) \propto \text{const}$ or $\pi(\gamma_{0,1}) \propto \text{const}$, $\pi(\gamma_{0,2}) \propto \text{const}$. The first order random walk prior controls abrupt jumps in the differences $\gamma_{0,m} - \gamma_{0,m-1}$, while the second order random walk prior penalizes deviations from a linear trend. The variance parameter δ_0^2 controls the amount of the penalization and acts as a smoothness parameter. The smaller the variance parameter, the stronger is the penalization. The joint prior for the parameter $\boldsymbol{\gamma}_0$ as the product of the conditional densities is

$$\pi\left(\boldsymbol{\gamma}_0 \mid \delta_0^2\right) \propto \left(\frac{1}{\delta_0^2}\right)^{\frac{k_0}{2}} \exp\left(-\frac{1}{2\delta_0^2}\boldsymbol{\gamma}'_0\mathbf{K}_0\boldsymbol{\gamma}_0\right), \tag{10.5}$$

with penalty matrix $\mathbf{K}_0$ of the form $\mathbf{K}_0 = \mathbf{D}'\mathbf{D}$, where $\mathbf{D}$ is a first or second order difference matrix and $k_0 = \text{rank}(\mathbf{K}_0)$. A standard prior option for the variance parameter is a diffuse inverse gamma prior $\delta_0^2 \sim \text{IG}(a_0, b_0)$ with small $a_0 > 0$, $b_0 > 0$. It is also possible to choose an improper gamma-type prior in the case when $a_0 \le 0$, $b_0 \le 0$ or especially $a_0 \le 0$, $b_0 = 0$, for example $a_0 = -0.5$, $b_0 = 0$, corresponding to a flat prior $\pi(\delta_0) \propto \text{const}$ for the standard deviation δ_0.

In the more general case of the extended model (10.3), unknown nonparametric functions $f_j(z_j)$ of continuous covariates are modelled through Bayesian P-splines as well. Then the vector $\boldsymbol{\eta}$ of predictors $\eta_i = \eta_i(t_i)$ evaluated at observed lifetimes t_i, $i = 1, \dots, n$, can always be represented, after reindexing, as a high-dimensional predictor in generic matrix notation

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_0\boldsymbol{\gamma}_0 + \cdots + \mathbf{Z}_q\boldsymbol{\gamma}_q.$$

The design matrix $\mathbf{X}$ has rows $\mathbf{x}'_i$ and the design matrices $\mathbf{Z}_0, \dots, \mathbf{Z}_q$ are constructed from basis functions representing the functions g_0, $f_1, \dots, f_q$ and $\boldsymbol{\gamma}_0, \dots, \boldsymbol{\gamma}_q$ are corresponding vectors of basis coefficients, as described above. The general form of smoothness priors for $\boldsymbol{\gamma}_j$ is once again of form (10.5) with variance parameter δ_j^2, precision or penalty matrix $\mathbf{K}_j$ and $k_j = \text{rank}(\mathbf{K}_j)$, $j = 0, 1, \dots, q$. For further details and extensions to time-varying, spatial and group-specific effects see Hennerfeind et al. (2006).

10.4 Posterior Inference

To illustrate posterior inference, we use the full likelihood together with the predictor $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_0\boldsymbol{\gamma}_0$. The posterior has the general form

$$\begin{aligned}&\pi\big(\boldsymbol{\beta}, \boldsymbol{\gamma}_0, \delta_0^2, \tau_1^2, \dots, \tau_p^2, \varphi\big)\\ &\quad\propto L(\boldsymbol{\beta}, \boldsymbol{\gamma}_0)\pi\big(\boldsymbol{\gamma}_0 \mid \delta_0^2\big)\pi\big(\delta_0^2\big)\prod_{j=1}^{p}\pi\big(\beta_j \mid \tau_j^2\big)\pi\big(\tau_1^2, \dots, \tau_p^2, \varphi\big),\end{aligned}$$

where $\pi(\tau_1^2, \dots, \tau_p^2, \varphi)$ is a generic notation for the priors for the variance parameters τ_j^2 the and other parameters or random variables φ defined through the hierarchical formulation of shrinkage priors in Sect. 10.3. Thus, for all shrinkage priors, the full conditional of the regression parameters β is

$$\pi\big(\boldsymbol{\beta} \mid \cdot\big) \propto \exp\left\{ l(\boldsymbol{\beta}, \boldsymbol{\gamma}_0) - \frac{1}{2}\boldsymbol{\beta}'\mathbf{D}_\tau^{-1}\boldsymbol{\beta} \right\},$$

where $\mathbf{D}_\tau = \text{diag}(\tau_1^2, \dots, \tau_p^2)$ denotes the diagonal matrix of the variance parameters. This distribution has no closed form to draw a new proposed state for the Markov chain. We use an MH-algorithm with IWLS-proposal to update the regression coefficients based on a second order Taylor expansion of the log-likelihood $l(\boldsymbol{\beta}, \boldsymbol{\gamma}_0)$ at the current state $\boldsymbol{\beta}^{(c)}$ of the parameter vector to construct a multivariate Gaussian proposal. MH-steps for updating the coefficients $\boldsymbol{\gamma}_0$ can conceptually be carried out in the same way as for $\boldsymbol{\beta}$, replacing the (conditional) precision matrix $\mathbf{D}_\tau^{-1}$ by the precision matrix of the (conditional) Gaussian prior $\mathbf{K}/\delta_0^2$.

The full conditionals for corresponding shrinkage components like the variance parameters τ_j^2, or ψ_j^2 $j = 1, \dots, p$, the shrinkage parameter λ or mixing parameter ω are all known densities so that the updates for the Markov chain are available via Gibbs steps. More details are available in Sect. 10.3 of the electronic supplement.

For the extended model (10.3), additional MH steps based on IWLS proposals are required to update the basis coefficients $\boldsymbol{\gamma}_j$ of the functions $f_j(z_j)$. If we rely on posterior inference based on the partial likelihood, we replace the log-likelihood $l(\boldsymbol{\beta}, \boldsymbol{\gamma}_0)$ through the partial log-likelihood $\text{pl}(\boldsymbol{\beta})$, compare again the electronic supplement.

10.5 Simulations

Here we report only a selection of results; additional material, in particular for inference based on the partial likelihood, is contained in the electronic supplement. The analysis of the parametric and nonparametric models with P-spline and Weibull baseline was done with the free BayesX software available from http://www.bayesx.org.

We measure the estimation accuracy based on the mean squared errors (MSE) over $r = 50$ runs with sample size $n = 200$ if only linear effects are modeled and $n = 1000$ if nonlinear effects are included in the predictor. Let $\mathbf{V} = (\mathbf{X}'\mathbf{X})^{-1}/n$ be the empirical covariance matrix of the regressors, then the MSE of $\hat{\boldsymbol{\beta}}$ in the r-th simulation replication is given by $\text{MSE}_r(\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta})'\mathbf{V}(\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta})$. For the log-baseline g_0 we get $\text{MSE}_r(\mathbf{g}_0) = (\hat{\mathbf{g}}_{0,r} - \mathbf{g}_0)'(\hat{\mathbf{g}}_{0,r} - \mathbf{g}_0)/n$, where $\hat{\mathbf{g}}_{0,r} = (\hat{g}_{0,r}(t_1), \ldots, \hat{g}_{0,r}(t_n))'$ denotes the vector of function evaluations of the log-baseline $\hat{g}_0$ in the r-th simulation. To compare the results of P-spline-based log-baseline estimation and the corresponding Breslow estimates from the partial likelihood approaches, we use the trapezoidal rule to compute the cumulative baseline hazard and the corresponding formula for the MSE_r. Concordantly, if f denotes the vector of nonlinear effects of covariate x with estimate $\hat{\mathbf{f}}_r = (\hat{f}(x_1), \ldots, \hat{f}(x_n))'$, the MSE is given as $\text{MSE}_r(\mathbf{f}) = (\hat{\mathbf{f}}_r - \mathbf{f})'(\hat{\mathbf{f}}_r - \mathbf{f})/n$. Additionally, we report the average number of correct and incorrect zero coefficients in the final models achieved after applying one of the hard shrinkage rules discussed in Sect. 10.3.4.

The simulation settings are extensions of those used in Hennerfeind et al. (2006). We consider $r = 50$ datasets with a sample size of $n = 1000$ life times. The covariates $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,9})'$ corresponding to the linear effects are randomly drawn from a multivariate Gaussian distribution with zero mean and covariance matrix chosen such that the correlation between $\mathbf{x}_j$ and $\mathbf{x}_k$ is $\text{corr}(x_{i,j}, x_{i,k}) = \rho^{|j-k|}$ with $\rho = 0.5$ and the variances are set to $0, 25^2$ to get survival times in the area where the baseline changes (model B below). The covariate x_{10} corresponding to the nonlinear effect was independently drawn from a uniform $U[-3, 3]$ distribution. The lifetimes are generated via the inversion method (Bender et al. 2005) from the model

$$\boldsymbol{\beta} = (-0.7, -0.7, 0, 0, 0, -0.7, 0, 0, 0)',$$
$$\lambda(t) = \lambda_0(t) \exp\bigl(\boldsymbol{x}'\boldsymbol{\beta} + \sin(x_{10})\bigr).$$

The nonlinear effect is centered due to identification arguments leading to an intercept term in the predictor. To model more flexible baseline hazards, a linear but non-Weibull baseline hazard of the form A and a bathtub-shaped baseline hazard of the form B have been chosen:

$$\text{Model A:} \quad \lambda_0(t) = 0.25 + 2t,$$
$$\text{Model B:} \quad \lambda_0(t) = \begin{cases} 0.75(\cos(t) + 1.5), & t \leq 2\pi, \\ 0.75(1 + 1.5), & t > 2\pi. \end{cases}$$

The latter assumes an initially high baseline risk that decreases after some time and increases again later on until time $t = 2\pi$ from where the hazard stays constant.

Censoring times are generated in two steps. First, a random proportion of 17 % of the generated observations T_i is assigned to be censored. Then in the second step the censoring times for this random selection are drawn from the corresponding uniform distributions $U[0, T_i]$. The hyperparameters of the Bayesian lasso are set to the weakly informative values $a_\lambda = b_\lambda = 0.01$ and the hyperparameters of the Bayesian NMIG are $\nu_1 = 1$, $\nu_0 = 0.000025$, $a_\psi = 5$ and $b_\psi = 25$. These values are chosen to assign a marginal prior probability of about 0.8 to fall into the interval $[-2, 2]$ to each regression coefficient. Further we use 30000 iterations with a burnin of 10000 and thin the chain by 20 which results in an MCMC sample of size 1000. Before we describe our results, we introduce some abbreviations to reduce the writing.

B, BT, BL, BN Bayesian models based on the full likelihood with P-spline baseline hazard without penalization (B), with the predictor that contains only the nonzero effects (BT), with lasso (BL) and NMIG (BN) prior.

For the Bayesian approaches, the hard shrinkage methods described in Sect. 10.3.4 are additionally assigned with the following.

HS.CI95, HS.STD if hard shrinkage is done via the 95 % credible region (HS.CI95) or the one standard error region (HS.STD),

HS.IND if hard shrinkage for NMIG is done via indicator variables.

For example, PL.BN-HS.IND denotes the Bayesian partial likelihood model under NMIG penalty when the covariate specific indicators are used to select the covariates for the final model.

We summarize the main results for the models A and B. The results for further 3 simulation models are provided with additional results using a Weibull model for the baseline in the electronic supplement together with mentioned and not displayed results for the models A and B. In contrast to the models in the electronic supplement, we restrict here the analysis to the Bayesian methods based on the full likelihood with P-spline approximation for the baseline. At present, there are no distributed packages in R available to perform frequentist lasso regression in combination with nonlinear effects for Cox models. Ridge regression is possible but the shrinkage parameter lambda has to be prespecified.

In Fig. 10.2, the MSEs of the estimated regression coefficients of simulation model A are shown together with the MSEs if the hard shrinkage criteria of Sect. 10.3.4 are applied. As in model 1, the Bayesian NMIG performs better than the Bayesian lasso regardless of whether hard shrinkage is applied or not and the MSEs are very similar to the MSE of BT, where only the true nonzero effects are included in the predictor. The box plots of two selected estimated coefficients for model B with bathtub shaped baseline are shown in Fig. 10.3. The box plots corresponding to the different methods are very similar except those of the Bayesian NMIG for the zero coefficients which show a higher concentration around zero. Thus, the hard shrinkage rules for the Bayesian NMIG are leading to comparable results as shown in Fig. 10.2, since the non-influential covariates are assigned an effect very close to zero anyway, i.e., it is negligible if they are removed from the final model.

The variable selection feature of the Bayesian NMIG is highlighted in Fig. 10.4 where the box plots of the relative frequencies of the indicator variables $\nu_1 = 1$

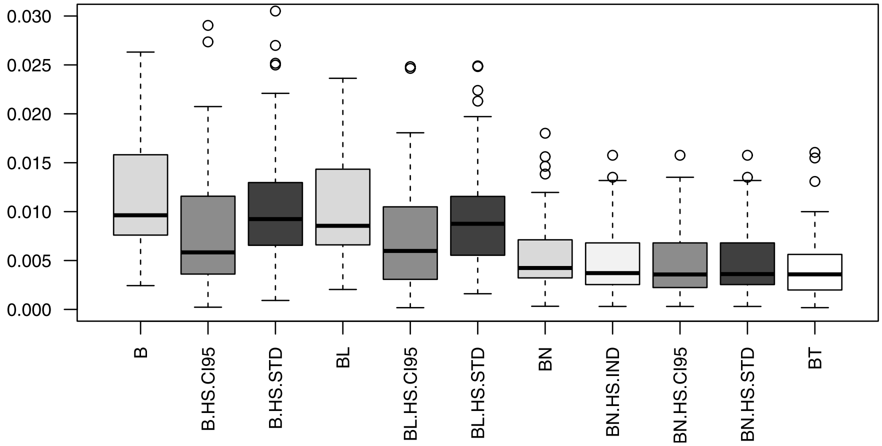


Fig. 10.2 Box plots of the mean squared errors for the regression coefficients of simulation model A in combination with the mean squared errors if hard shrinkage is used

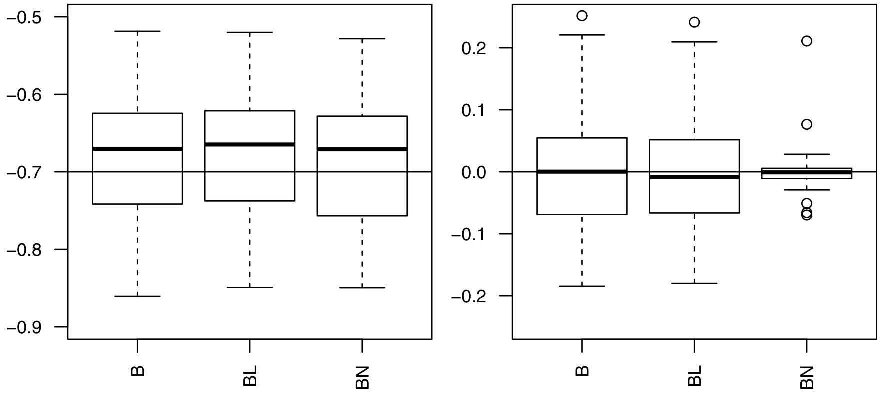


Fig. 10.3 Box plots of two of the estimated coefficients for the simulation under simulation model B (*left panel* $\hat{\beta}_2$ and *right panel* $\hat{\beta}_3$). *The black horizontal lines* mark the values of the true linear effects

are shown for the more complex model B. The relative frequencies of the nonzero effects are nearly one with very small standard deviation. For the zero effects, the relative frequencies are shifted towards zero and clearly fall below the selection cut off value of 0.5 so that the relative frequencies of the indicator variables seem to provide a good resource to select the important covariates. Similar results arise in model 1 and A, compare the electronic supplement.

In Fig. 10.5, we see the estimates of the baseline (Fig. 10.5, left side) and the nonlinear effect (Fig. 10.5, right side) for one selected dataset if the Bayesian NMIG is applied to model B. The vertical lines at the time of the baseline axis mark the observed events. In the interval [0, 6], where most of the observations occur, the

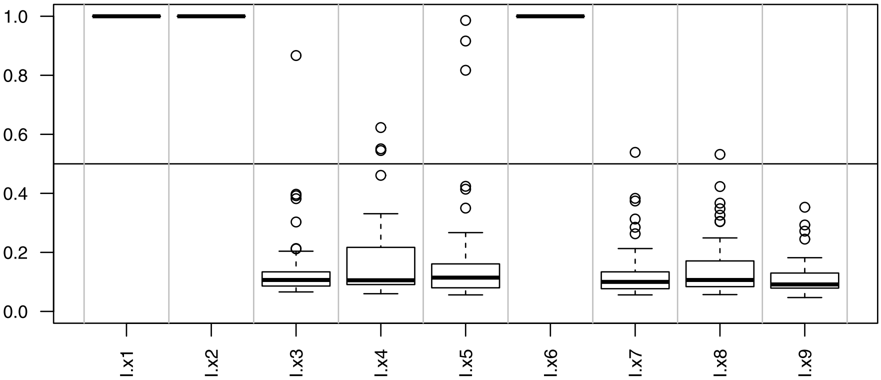


Fig. 10.4 Box plots of the relative frequencies of the indicator variables for NMIG penalty in simulation model B

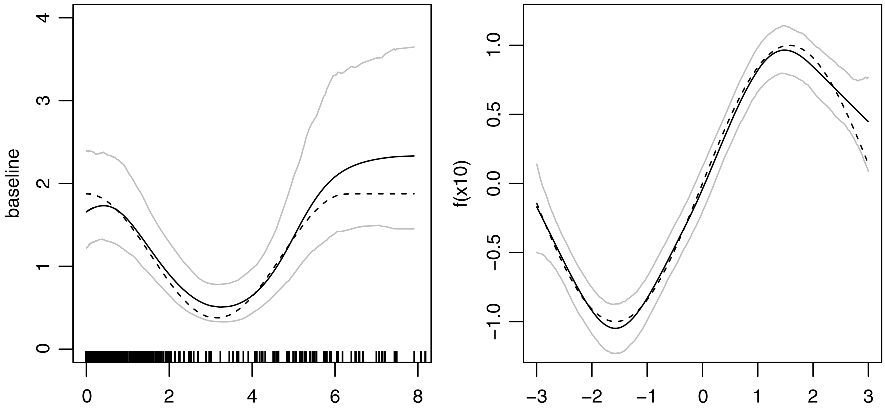


Fig. 10.5 Baseline hazard estimation (*left side, solid black line*) and estimation of the nonlinear effect $f(x) = \sin(x)$ (*right side, solid black line*) under the Bayesian NMIG penalty for one selected dataset under simulation model B together with the 2.5 % and 97.5 % empirical quantiles (*grey lines*). *The black dashed lines* indicate the corresponding true baseline or true nonlinear effect

P-spline baseline approximates the true baseline very well. The deviations get larger, the less observations are available when time increases, which results in an increasing MSE. The results for baseline estimation are as good as in model A and the results of the different methods are also comparable to each other. The same holds for the estimation of the nonlinear effect, compare the electronic supplement.

The frequencies of the selected final models and the number of true estimated coefficients are listed for the different hard shrinkage rules in Table 10.1. The best case of 50 is reached if HS.CI95 is applied, which is here not surprising because the Bayesian NMIG behaves very similar to BT with asymptotic normal estimates. Additionally, the second and third column in Table 10.1 show the average number of the true estimated nonzero coefficients and true estimated zero coefficients for

Table 10.1 Number of "true" estimated coefficients where $\hat{\beta} \neq 0, \beta \neq 0$ denotes the case that the estimated effect is nonzero ($\hat{\beta} \neq 0$) when the corresponding true effect is nonzero ($\beta \neq 0$) and $\hat{\beta} = 0, \beta = 0$ denotes the case that the estimated effect is zero ($\hat{\beta} = 0$) when the corresponding true effect is zero ($\beta = 0$). *The columns (MF)* display the frequencies of the final models that contain only the three effects $\beta_1 \neq 0, \beta_2 \neq 0, \beta_6 \neq 0$ for model A and B

	Model A			Model B		
	$\hat{\beta} \neq 0$	$\hat{\beta} = 0$	MF	$\hat{\beta} \neq 0$	$\hat{\beta} = 0$	MF
	$\beta \neq 0$	$\beta = 0$		$\beta \neq 0$	$\beta = 0$	
BEST	3	6	50	3	6	50
B.HS-STD	3	4.26	7	3	3.94	4
BL.HS-STD	3	4.46	8	3	4.06	5
BN.HS-STD	3	5.94	47	3	5.90	45
B.HS-CI95	3	5.66	37	3	5.62	35
BL.HS-CI95	3	5.72	38	3	5.70	37
BN.HS-CI95	3	6	50	3	5.96	48
BN.HS-IND	3	5.92	46	3	5.82	43

the 50 datasets. All hard shrinkage methods reach the optimal value of three for the true nonzero coefficients. The highest values for the true zero coefficients are again achieved for the Bayesian NMIG.

10.6 Application to AML Data

To illustrate the presented methods and to compare them with frequentist alternatives, we analyze data for patients diseased with cytogenetically normal acute myeloid leukemia (CN-AML). AML is a cancer of the myeloid line of blood cells which is characterized by the rapid growth of abnormal white blood cells that accumulate in the bone marrow and interfere with the production of normal blood cells. Gene expression profiling can be used to develop a gene signature that predicts the overall survival time of patients in combination with prognostic factors like molecular markers and patient characteristics.

We use data provided by U. Mansmann (LMU München), described in detail and analyzed in Metzeler et al. (2008) and Benner et al. (2010). There are two independent cohorts of patients available. The training cohort stems from the multicenter AMLCG-199 trial of the German AML Cooperative Group between 1999 and 2003 and consists of 163 adult patients with CN-AML, where 35.0 % of the observed survival times are censored. In the training data, the median survival time is 280 days with range from 0 to 2399 days. The independent test cohort consists of 80 patients who were diagnosed with CN-AML in 2004. In the test data we have a median survival time of 247.5 days with range 1 to 837 days and 57.5 % of censored survival times. In both cohorts, survival time is defined as time from study entry until death from any cause. The original data comprises 44754 microarray probe sets for each individual. Univariate Cox scores, measuring the correlation between each of the probe sets and the survival time in the training cohort, are used to rank and to reduce the number of probe sets. We present results based on the probe sets with the 50 highest ranks; additional results for the first 200 probe sets are given in the electronic supplement. As additional prognostic covariates, we include the age (AGE) of the patient and the two molecular markers FLT3 (tandem duplications of the fms-like tyrosine kinase 3) and NPM1 (mutations in the nucleophosmin 1).

We compare the Bayesian ridge, lasso and the NMIG prior approach based on the partial and the full likelihood with some frequentist methods, in particular a backward-stepwise procedure based on the AIC criterion under Cox's proportional hazards model (STEP) and the frequentist lasso (Pen.L) and ridge regression (Pen.R) as provided in the R package {penalized} (Goeman 2010). In the latter case, the penalization parameters are determined by n-fold generalized cross validation. To illustrate simultaneous shrinkage and smoothing, we model the log-baseline hazard rate and AGE by P-splines of degree 3, with 20 knots and a random walk penalty of order 2. For MCMC runs, we use 30000 iterations with a burnin of 15000 and thin the chain by 15, which results in an MCMC sample of size 1000. The hyperparameters of the Bayesian lasso, ridge and NMIG are the same as in the simulations settings of Sect. 10.5. To avoid manual tuning of the regularization priors, the continuous covariates in the training and test data were standardized to have zero mean and unit variance.

To measure predictive accuracy, the time-dependent empirical Brier score $\mathrm{BS}(t)$ as proposed by Graf et al. (1999) is used, more specifically the integrated Brier score (IBS)

$$\mathrm{IBS} = \frac{1}{t^*} \int_0^{t^*} \mathrm{BS}(s)\, ds,$$

which can further be used to derive a measure of explained variation $R^2_{\mathrm{IBS}} := 1 - \mathrm{IBS}/\mathrm{IBS}_0$ with IBS_0 defined as the integrated Brier score corresponding to the Kaplan-Meier (KM) estimate of the survival function. Cumulative baseline hazards are computed via the Breslow estimator.

The IBSs and the corresponding measures of explained variation for a selection of the estimated models can be found in Table 10.2 for the case of 50 preselected probe sets (complete results as well as results for 200 probe sets are available in the electronic supplement). Overall, the smallest IBS is achieved by Bayesian ridge in combination with partial likelihood estimation (PL.BR, $\mathrm{IBS} = 0.168$ in the test set), followed closely by the Bayesian lasso based on the partial likelihood (PL.BL, $\mathrm{IBS} = 0.1701$). In general, the IBSs seem to suggest that the best strategy to achieve precise predictions is to include all covariates without hard shrinkage but to apply regularization to the coefficient vector. This claim is further supported by the result for the NMIG prior structure (that enables selection/deselection of covariates based on the latent binary indicator) that leads to somewhat deteriorated IBS. The same result holds for the models where a hard threshold rule is applied after a first estimation run based on the standard deviation or the 95 % credible interval.

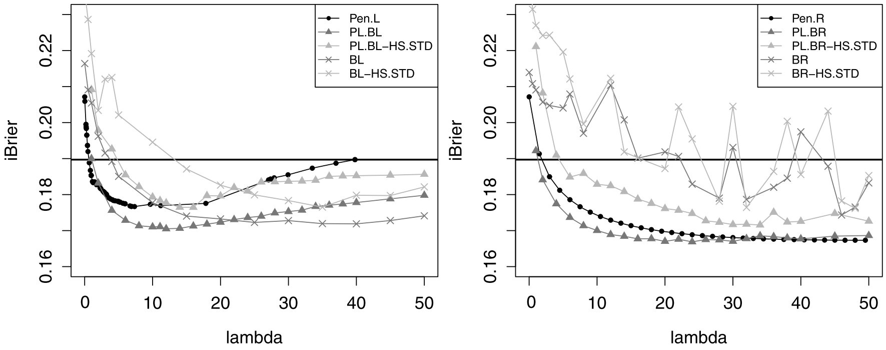


Fig. 10.6 Paths of the integrated Brier scores for varying complexity parameter with linear effect of AGE in case of lasso type regularization (*left panel*) and ridge type regularization (*right panel*). *The horizontal line* marks the IBS 0.1897 of the Cox proportional hazards model with the 3 pheno covariates

Figure 10.6 illustrates the impact of the regularization parameter on the IBS for the different lasso versions (left panel) and the ridge type models (right panel). The first observation that we can make from these figures is that again the Bayesian ridge based on the partial likelihood leads to the smallest IBS and is also fairly insensitive with respect to the regularization parameter. In addition, the behavior of the frequentist ridge is quite close to the Bayesian version. For ridge regression in combination with the full likelihood, there seems to be some instability in estimation that yields abrupt changes in the IBS even for small variations of the regularization parameter. In general, the lasso variants are in closer agreement and also do not show this irregular behavior. Again, the Bayesian lasso based on the partial likelihood performs remarkably well, but full likelihood based estimates are close and even yield improved performance for large regularization parameters.

Figure 10.7 shows the paths of estimated regression coefficients as a function of the complexity parameter for four covariates associated with Cox score ranks 1, 11, 12, and 21 for four different estimation approaches. It turns out that especially the probe sets associated with Cox score ranks 11 and 12 yield large estimates and are therefore deemed important over wide ranges of the values of the complexity parameter. In particular, in case of the Bayesian ridge, estimated effects are pretty much insensitive with respect to the values of the complexity parameter.

Figure 10.8 shows similar information for estimation results achieved with the NMIG prior structure. The left panel shows the estimated inclusion probabilities, i.e., the relative frequencies of the binary indicator differentiating between selected and deselected covariates. In this panel, the covariates with Cox score ranks 11 and 12 again stand out with large inclusion probabilities, clearly exceeding the threshold of 0.5. The right panel shows the inclusion probability as a function of the complexity parameter. In contrast to the probe sets with Cox score ranks 1 and 21 it turns out, that the inclusion probability for the probe sets with Cox score ranks 11 and 12 does not vary very much and always yields the conclusion that these probe sets should be contained in the model.

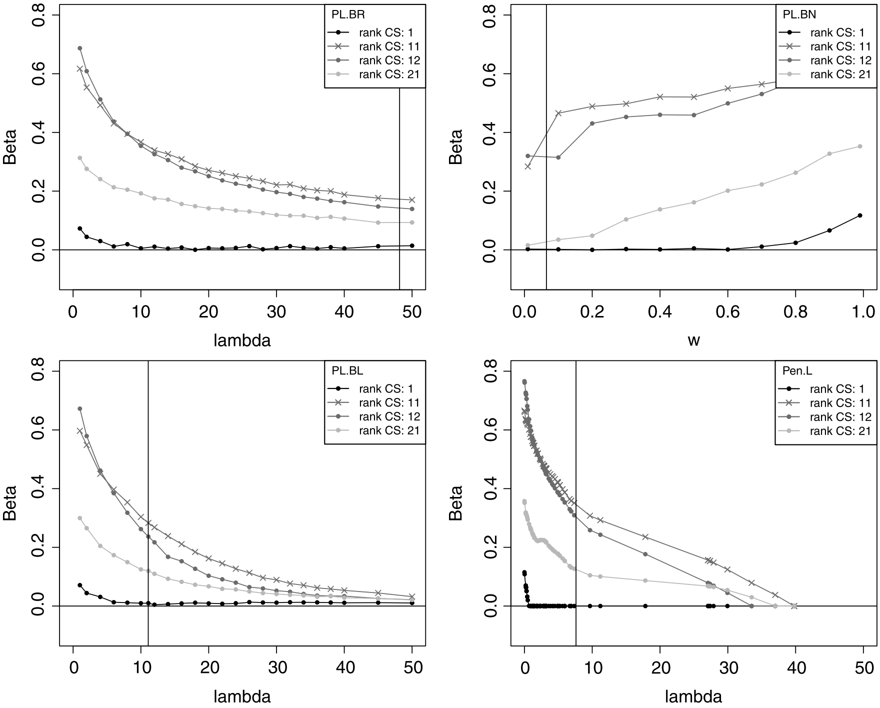


Fig. 10.7 Paths of four selected estimated coefficients (identified by the rank of the Cox score) as a function of the complexity parameter with linear effect of the AGE. *The vertical black solid line* marks the estimated values of the corresponding complexity parameters

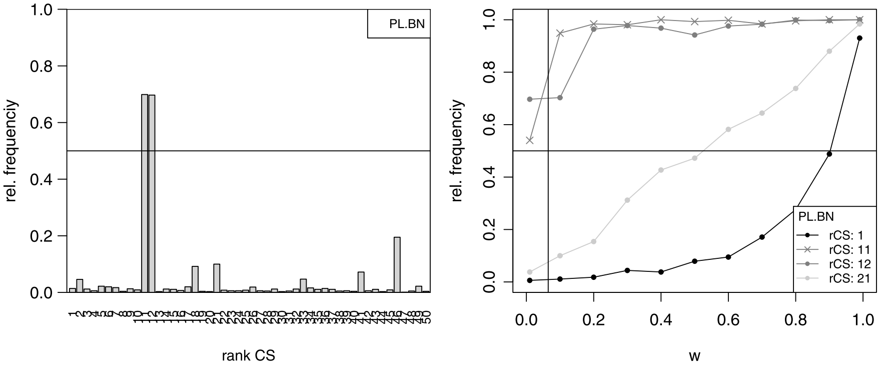


Fig. 10.8 *Left panel*: Estimated posterior relative frequencies of the of the Bayesian NMIG indicator variable value ν_1 with linear effect of the AGE. The probe sets are sorted according to the rank of the Cox scores that are displayed at the x axis. *Right panel*: Relative frequencies of four selected indicator variables as function of the complexity parameter ω. *The vertical black solid line* marks the estimated value $\hat{\omega}$

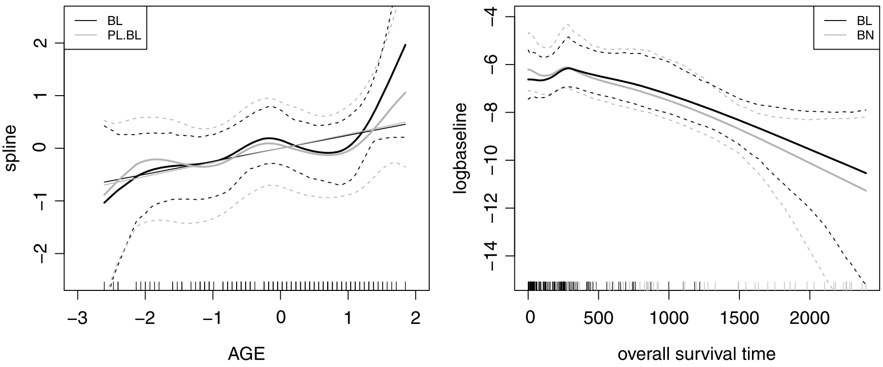


Fig. 10.9 *Left panel*: Nonlinear effect of the covariate AGE. *The solid bold lines* show the posterior mean estimates and *the dotted lines* mark the corresponding 95 % pointwise credible bands. *The straight lines* display the linear effect of the covariate AGE. *The black stripes* at the x axis mark the observed values. *Right panel*: Estimated log baseline hazard based on the posterior mean (*solid lines*) with the corresponding 95 % pointwise credible bands (*dashed lines*)

Finally, the left panel of Fig. 10.9 shows the estimated nonlinear effect of the covariate AGE and also includes the corresponding estimate from a linear model for comparison. While the spline estimate does show some nonlinearity, the associated credible intervals all cover the linear effect such that there is only weak evidence for the necessity of a nonlinear AGE effect. The right panel of Fig. 10.9 shows the estimated log-baseline hazard rate obtained with the full likelihood. After a short period of constant or moderately increasing baseline hazard, the hazard rate shows an almost linear decrease.

In summary, our analyses allow to conclude that, depending on the specific purpose of the analysis, different variants of Bayesian regularization seem to be more or less suitable. If the ultimate goal is good prediction, a full model combined with either ridge or lasso regularization is recommended with a small preference for the Bayesian ridge. If an easily understandable model is desired that can also easily be communicated to physicians or patients, regularization including the selection of effects, as for example available with the NMIG prior structure, will usually be preferable. In our analyses, we found strong evidence for the importance of the two probe sets associated with Cox score ranks 11 and 12. In addition, our model class allows us to validate the assumption of linearity of pheno covariates that are often available in addition to genetic information. While we did not find evidence for such a nonlinear effect in case of age, the ability to check for nonlinearity is still a significant improvement from an applied perspective.

10.7 Conclusions

We have developed different types of regularization priors for flexible hazard regression models that allow us to combine modelling of complex predictor structures

Table 10.2 IBS and R^2_{IBS} in the training and test data that uses the 50 probe sets with highest Cox scores. The pheno covariates FLT3, NPM1 and the AGE are always included in the models. CoxPH and CoxPH3 denotes the results from frequentist Cox models where the latter denotes the model which includes only the 3 pheno covariates. Bayesian models based on partial likelihood are labeled with the prefix PL. In particular PL.B denotes the Bayesian model without penalization and PL.BL uses the lasso and PL.BN the NMIG penalty. SUMps displays the number of probe sets and SUMp the number of pheno covariates in the model

	SUMps	SUMp	IBS.train	R2.train	IBS.test	R2.test
KM	0	0	0.2120	0	0.2055	0
CoxPH3	0	3	0.1713	0.1921	0.1897	0.0767
CoxPH	50	3	0.0995	0.5306	0.2081	−0.0126
STEP	16	3	0.1129	0.4674	0.1952	0.0499
Pen.L	11	3	0.1281	0.3957	0.1767	0.1402
Pen.R	50	3	0.1384	0.3470	0.1710	0.1680
PL.B	50	3	0.0989	0.5337	0.2120	−0.0318
PL.BR	50	3	0.1305	0.3847	0.1680	0.1824
PL.BL	50	3	0.1253	0.4090	0.1701	0.1719
PL.BN	50	3	0.1352	0.3624	0.1843	0.1030
PL.BR-HS.STD	6	3	0.1405	0.3373	0.1726	0.1599
PL.BR-HS.CI95	1	3	0.1648	0.2225	0.1836	0.1063
PL.BL-HS.STD	5	3	0.1372	0.3526	0.1783	0.1321
PL.BL-HS.CI95	1	3	0.1609	0.2410	0.1856	0.0966
PL.BN-HS.IND	2	3	0.1423	0.3289	0.1878	0.0861
B	50	3	0.1007	0.5252	0.2145	−0.0439
BR	50	3	0.1004	0.5266	0.1980	0.0363
BL	50	3	0.1027	0.5158	0.1876	0.0867
BN	50	3	0.1187	0.4403	0.1870	0.0897
BR-HS.STD	20	3	0.1078	0.4917	0.2026	0.0141
BR-HS.CI95	5	3	0.1365	0.3562	0.2210	−0.0755
BL-HS.STD	13	3	0.1241	0.4145	0.2104	−0.0238
BL-HS.CI95	3	3	0.1342	0.3669	0.2084	−0.0144
BN-HS.IND	2	3	0.1405	0.3372	0.2017	0.0185

with regularization of effects of possibly high-dimensional ($n < p$) covariate vectors. The basic advantages of the Bayesian regularization approach are two-fold: On the one hand, complex models can be built from blocks considered in previous approaches due to the modularity of MCMC simulations. On the other hand, the Bayesian formulation allows for the simultaneous estimation of all parameters involved while allowing for significance and uncertainty statements even about complex functions of these parameters. The restriction that posterior mean estimates in regularized regression models do not directly provide the variable selection property known for example from the frequentist lasso can be overcome by the latent indicator approach in the NMIG prior model. In the context of gene expression data, for example, the advantages of the Bayesian approach will be particularly valuable since flexible modelling of clinical covariates can be combined with regularization of microarray features. We will also investigate adaptive versions of the proposed regularization priors where separate smoothness parameters are added to the scale mixture representation. This should allow overcoming the necessity to standardize covariates up-front, since the priors are allowed to adapt to the varying scaling. The class of regularized regression models for survival times will be broadened by considering accelerated failure time (AFT) models based on imputation of censored observations as a part of the MCMC scheme.

Acknowledgements We thank Ulrich Mansmann for providing the CN-AML data and stimulating discussions concerning our results. Financial support by the German Research Foundation (DFG), grant FA 128/5-1/2 is gratefully acknowledged.

References

Bender, R., Augustin, T., & Blettner, M. (2005). Simulating survival times for Cox regression models. *Statistics in Medicine*, *24*, 1713–1723.

Benner, A., Zucknick, M., Hielscher, T., Ittrich, C., & Mansmann, U. (2010). High-dimensional Cox models: the choice of penalty as part of the model building process. *Biometrical Journal*, *52*, 50–69.

Brezger, A., & Lang, S. (2006). Generalized additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis*, *50*, 967–991.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society. Series B. Methodological*, *34*, 187–220.

Fahrmeir, L., Kneib, T., & Konrath, S. (2010). Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, *20*, 203–219.

Fan, J., & Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, *30*, 74–99.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*, 881–889.

George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, *7*, 339–373.

Goeman, J. J. (2010). L1-penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, *52*, 70–84.

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, *18*, 2529–2545.

Griffin, J. E., & Brown, P. J. (2005). *Alternative prior distributions for variable selection with very many more variables than observations*. Technical report, University of Warwick, Department of Statistics.

Griffin, J. E., & Brown, P. J. (2010). *Bayesian adaptive lassos with non-convex penalization*. Technical report, University of Warwick, Department of Statistics.

Hennerfeind, A., Brezger, A., & Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association*, *101*, 1065–1075.

Ishwaran, H., & Rao, S. J. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association*, *98*, 438–455.

Ishwaran, H., & Rao, S. J. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, *33*, 730–773.

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data* (2nd ed.). New York: Springer.

Kneib, T., Konrath, S., & Fahrmeir, L. (2010). High-dimensional structured additive regression models: Bayesian regularisation, smoothing and predictive performance. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *60*, 51–70.

Lang, S., & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, *13*, 183–212.

Li, Q., & Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis*, *5*, 847–866.

Lin, D. Y. (2007). On the Breslow estimator. *Lifetime Data Analysis*, *13*, 471–480.

Metzeler, K. H., Hummel, M., Bloomfield, C. D., Spiekermann, K., Braess, J., et al. (2008). An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood*, *112*, 4193–4201.

Panagiotelis, A., & Smith, M. (2008). Bayesian identification, selection and estimation of semiparametric functions in high-dimensional additive models. *Journal of Econometrics*, *143*, 291–316.

Park, M. Y., & Hastie, T. (2007). L1 regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *69*, 659–677.

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*, 681–686.

Scheipl, F. (2011). *Bayesian regularization and model choice in structured additive regression*. Dr. Hut Verlag.

Scheipl, F., Fahrmeir, L., & Kneib, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*. doi:10.1080/01621459.2012.737742

Sinha, D., Ibrahim, J. G., & Chen, M. H. (2003). A Bayesian justification of Cox's partial likelihood. *Biometrika*, *90*, 629–641.

Smith, M., & Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, *75*, 317–343.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, *16*, 385–395.

van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., Van't Veer, L. J., & Wessels, L. F. A. (2006). Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine*, *25*, 3201–3216.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.

Chapter 11
Robust Change Point Analysis

Marie Hušková

11.1 Introduction

The change point problem is usually treated by statistical procedures for detection of instabilities in statistical models. The problem is usually formulated in terms of hypothesis testing and estimation problem. Typically, we have observations $X_1, \ldots, X_n$ obtained at ordered time points and the basic task is to decide whether the model remains stable during the whole observational period or whether the model changes at some unknown point(s) or become generally instable. In case of change(s) in the model being detected, the further task is also to estimate the time of change and other parameters of the model in the periods where the model is stable. The former problem is formulated in terms of hypothesis testing whether the model remained stable during the observational procedures (null hypothesis) against an alternative that the model changes at least once. The latter problem is to estimate the location of time points where the model changes they are called *unknown change points* and to estimate further parameters.

Such problems are also called disorder problems or testing for presence of structural breaks (in econometrics) or testing for stability or segmented regression or switching regression in the regression setup.

If all n observations are available at the beginning of the statistical analysis, we speak about a *retrospective setup*. If observations are arriving sequentially and after each new observation, we have to decide whether the observations obtained so far indicate an instability or not we have a *sequential setup*.

Originally such problems were studied within statistical quality control, however nowadays there are many applications in various areas, e.g., medical research, econometrics, financial models, risk management, environmetrics, climatology. It brings a number of interesting theoretical problems.

M. Hušková (✉)
Department of Statistics, Charles University in Prague, Sokolovská 83, 186 75 Prague,
Czech Republic
e-mail: huskova@karlin.mff.cuni.cz

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_11, © Springer-Verlag Berlin Heidelberg 2013

There is a number of monographs and survey papers tackling the problem from various points of view, various generality, etc., e.g., Brodsky and Darkhovsky (1993), Basseville and Nikiforov (1993), Carlstein et al. (1994), Csörgő and Horváth (1997), Chen and Gupta (2000). Probably, the most influential papers in econometric literatures are papers by Andrews (1993), Bai and Perron (1998) and Perron (2006). Partial survey of basic procedures till 2000 can be found, e.g., in Antoch and Hušková (1999). Antoch et al. (2000) deal with various change point estimators including robust ones.

Statistical procedures (tests and estimators) for detection of a change were developed applying various principles. As a motivation can serve, the case of the known change point which leads to a variant of the two-sample problem in case of one change, while the k sample problem is related to more changes. Along this line one can develop maximum likelihood procedures, Bayesian procedures, nonparametric, robust ones, etc.

At first, the procedures for detection of a change were developed to detect instabilities in mean and variance in simpler models mostly for independent observations. Nowadays quite general models for instabilities appeared to be quite useful in applications, e.g., financial time series and econometrics. Quite popular and also practical are procedures for detection of change(s) in regression parameters in regression models and in time series. Typically, at first it is assumed that observations or/and the error terms are independent and identically distributed (i.i.d.) random variables with normal distribution and then it is checked whether slightly modified procedures can be also used under weaker assumptions.

The rest of the contribution is divided into three parts. Section 11.2 deals with M-tests for a change in linear models. Various M-test statistics are developed and their theoretical properties are presented. Section 11.3 concerns estimators of a single as well as multiple changes. Section 11.4 gives a short inside into the area of sequential M procedures and rank based procedures. Some open problems are formulated in Sect. 11.5.

11.2 *M*-Procedures for Detection of a Change in Regression

11.2.1 Formulation of the Problem and Procedures

The basic *regression model with a change after an unknown time point* k^* has the form:

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{x}_i^T \boldsymbol{\delta} I\{i > k^*\} + e_i, \quad i = 1 \ldots, n, \tag{11.1}$$

where $k^* = k_n^*(\leq n)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{\delta} = \boldsymbol{\delta}_n = (\delta_{1n}, \ldots, \delta_{pn})^T \neq \mathbf{0}$ are unknown parameters, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T, x_{i1} = 1, i = 1, \ldots, n$ are known regression vectors. Finally, $e_1, \ldots, e_n$ are random errors fulfilling regularity conditions discussed below. Function $I\{A\}$ denotes the indicator of the set A.

Model (11.1) describes the situation where the first k^* observations follow the linear model with the parameter $\boldsymbol{\beta}$ and the remaining $n - k^*$ observations follow the linear regression model with the parameter $\boldsymbol{\beta} + \boldsymbol{\delta}$. The parameter k^* is called *the change point*.

The main tasks are:

- to test the hypothesis of "no change" (H_0) versus an alternative "there is a change" (H_1), or in other words, to test

$$H_0 : k^* = n \quad \text{against } H_1 : k^* < n, \tag{11.2}$$

- in case of rejection H_0 to estimate location of the change point k^* and other parameters before and after the change.

Quite often, assuming additionally that the error terms are i.i.d. random variables the likelihood principle is applied both to obtain the test and estimators of the parameters and then we search whether eventually modified procedures can be applied under a weaker setup. Assuming $e_1, \ldots, e_n$ being i.i.d. with $N(0, \sigma^2), \sigma^2 > 0$ known, and under mild assumptions on the regression vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ the maximum likelihood principle leads to the test statistic (see, e.g., Csörgő and Horváth 1997)

$$T_{n,0} = \max_{p \le k < n-p} \left\{ \mathbf{S}_k^T \mathbf{C}_k^{-1} \mathbf{C}_n \left(\mathbf{C}_k^0\right)^{-1} \mathbf{S}_k \right\} / \sigma^2, \tag{11.3}$$

where $\mathbf{S}_k, k = 1, \ldots, n$ are partial sums of weighted L_2-residuals defined as

$$\mathbf{S}_k = \sum_{i=1}^{k} \mathbf{x}_i \left(Y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_n\right), \quad k = 1, \ldots, n, \tag{11.4}$$

$$\mathbf{C}_k = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T, \quad \mathbf{C}_k^0 = \mathbf{C}_n - \mathbf{C}_k. \tag{11.5}$$

Here $\hat{\boldsymbol{\beta}}_n$ is the least squares estimator of $\boldsymbol{\beta}$ based on $Y_1, \ldots, Y_n$. With a little algebra, it can be shown that $T_{n,0}$ can be expressed also as

$$T_{n,0} = \max_{p \le k < n-p} \left\{ \left(\hat{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_k^0\right)^T \left((\mathbf{C}_k)^{-1} + \left(\mathbf{C}_k^0\right)^{-1}\right)^{-1} \left(\hat{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_k^0\right) \right\} / \sigma^2, \tag{11.6}$$

where $\hat{\boldsymbol{\beta}}_k$ and $\hat{\boldsymbol{\beta}}_k^0$ are least squares estimators of $\boldsymbol{\beta}$ based on $Y_1, \ldots, Y_k$ and $Y_{k+1}, \ldots, Y_n$, respectively, i.e., our test statistic $T_{n,0}$ can be expressed as a functional of the differences $\hat{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_k^0, k = p, \ldots, n-p$ and it is clearly sensitive w.r.t. a change in regression parameters.

Large values of $T_{n,0}$ indicate that H_0 is violated. However under quite mild assumptions $T_{n,0} \to \infty$ in probability as $n \to \infty$ even under H_0, therefore other functionals of $\mathbf{S}_k = \sum_{i=1}^{k} \mathbf{x}_i (Y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_n), k = 1, \ldots, n$ bounded in probability and still sensitive w.r.t. a change in regression parameter are often used. If σ^2 is unknown (this is the usual case) we replace it by a suitable estimator. In the considered setup (model (11.1) with $N(0, \sigma^2)$ i.i.d. random errors), the maximum likelihood estimator $\hat{k}^*$ of the change point k^* is defined as

$$\hat{k}_n^* = \min\Big(p \le k \le n - p;\, \big(\mathbf{S}_k^T \mathbf{C}_k^{-1} \mathbf{C}_n \big(\mathbf{C}_k^0\big)^{-1} \mathbf{S}_k\big) \max_{p \le j < n-p} \big(\mathbf{S}_j^T \mathbf{C}_j^{-1} \mathbf{C}_n \big(\mathbf{C}_j^0\big)^{-1} \mathbf{S}_j\big)\Big), \tag{11.7}$$

that is a consistent estimator of the change point k^* under mild conditions. The above test and estimator including their modifications are expressible as functionals of L_2 (least squares) estimators therefore they are sometimes called L_2 procedures. These procedures can be applied even under milder assumptions on the distribution of the error terms, e.g., Csörgő and Horváth (1997).

It is well known that L_2 estimators and related test procedures are sensitive w.r.t. outliers and behave also quite poorly in case of errors with heavy tailed distributions, see Huber (1981) and Jurečková and Sen (1996) among others for usual regression setup without a change. The same applies to the above tests and estimators for detection of a change in model (11.1), i.e., applying them we can wrongly reject the null hypothesis (no change) due to the presence of an outlier or heavy tailed distribution of the error terms and $\hat{k}_n^*$ is wrongly classified as an estimator of change point. Therefore, it is desirable to develop test procedures that are sensitive w.r.t. a change in regression parameters but insensitive w.r.t. outliers. To develop such tests, we can modify the test statistic $T_{n,0}$ along the line of robust methods, i.e., the least squares estimators and L_2 residuals are replaced by their robust counter parts, in other words replace them by M-estimators and M-residuals, respectively.

There are procedures focusing on detecting outliers in regression models as well as in time series, for more information see, e.g., the contribution by Galeano and Peña (Chap. 15) where such procedures are developed and studied for various univariate as well as multivariate time series. Another possibility of identification of outliers is the conditional quantile approach, see the contribution by Barme, Chap. 3.

Recall the definition of M-estimators. In model (11.1) with $k^* = n$ the M-estimator $\hat{\boldsymbol{\beta}}_n(\psi)$ of the regression parameter $\boldsymbol{\beta}$ generated by a convex loss function ρ with the related monotone score function $\psi = \rho'$ a.s. is defined as a minimizer of

$$\sum_{i=1}^{n} \rho\big(Y_i - \mathbf{x}_i^T \mathbf{b}\big)$$

w.r.t. $\mathbf{b}$, which often reduces to finding the solution of the equation

$$\sum_{i=1}^{n} \mathbf{x}_i \psi\big(Y_i - \mathbf{x}_i^T \mathbf{b}\big) = \mathbf{0}.$$

For further information, see, e.g., Huber (1981), Jurečková and Sen (1996). Then the M-residuals are defined as

$$\hat{e}_{i,n}(\psi) = \psi\big(Y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_n(\psi)\big), \quad i = p, \dots, n - p, \tag{11.8}$$

and the related partial sums of weighted M-residuals as

$$\mathbf{S}_k(\psi) = \big(S_{k1}(\psi), \dots, S_{kp}(\psi)\big)^T = \sum_{i=1}^{k} \mathbf{x}_i \hat{e}_{i,n}(\psi), \quad k = p, \dots, n-p. \tag{11.9}$$

Various functionals of $\mathbf{S}_k(\psi), k = p, \dots, n-p$, can be used as robust test procedures. Particularly, we can introduce two robust analogs of $T_{n,0}$:

$$T_{n,0}(\psi) = \max_{p \le k < n-p} \big\{\mathbf{S}_k^T(\psi)\mathbf{C}_k^{-1}\mathbf{C}_n\big(\mathbf{C}_k^0\big)^{-1}\mathbf{S}_k(\psi)\big\}/\hat{\sigma}_n^2(\psi) \tag{11.10}$$

and

$$T_{n,00}(\psi)$$
$$= \frac{\max_{p\le k<n-p}\{(\hat{\boldsymbol{\beta}}_k(\psi) - \hat{\boldsymbol{\beta}}_k^0(\psi))^T((\mathbf{C}_k)^{-1} + (\mathbf{C}_k^0)^{-1})^{-1}(\hat{\boldsymbol{\beta}}_k(\psi) - \hat{\boldsymbol{\beta}}_k^0(\psi))\}}{\hat{\sigma}^2(\psi)}, \tag{11.11}$$

where $\hat{\sigma}_n^2(\psi)$ is a suitable standardization possibly depending on observations. Clearly, $T_{n,0}(\psi)$ and $T_{n,00}(\psi)$ are identical for $\psi(x) = x$ while they generally differ otherwise, but they are at least asymptotically equivalent. Since $T_{n,00}(\psi)$ is computationally more demanding $T_{n,0}(\psi)$ and its modifications are preferred.

Since $T_{n,0}(\psi) \to \infty$ in probability as $n \to \infty$ even under H_0 test statistics that are bounded in probability under H_0 and tend to ∞ under alternatives are preferred. For instance, the following statistics have the desired properties under mild conditions:

$$T_{n,1}(\psi) = \sup_{0<t<1} \left\{ \frac{\mathbf{S}_{\lfloor (n+1)t \rfloor}^T(\psi)\mathbf{C}_n^{-1}\mathbf{S}_{\lfloor (n+1)t \rfloor}(\psi)}{\hat{\sigma}_n^2(\psi)} \right\} \tag{11.12}$$

and

$$T_{n,B}(\psi) = \frac{1}{n}\sum_{k=p}^{n-p} \frac{\mathbf{S}_k^T(\psi)\mathbf{C}_n^{-1}\mathbf{S}_k(\psi)}{\hat{\sigma}_n^2(\psi)}, \tag{11.13}$$

where $\lfloor a \rfloor$ denotes the integer part of a. Statistic $T_{n,1}(\psi)$ is called *max-type* while $T_{n,B}(\psi)$ is *sum-type* or *Bayesian type*. Sometimes simple statistics depending only on $S_{k,1}, k = p, \dots, n-p$, that are functionals of partial sums of M-residuals are used. A number of other test statistics have been proposed and studied, e.g., Hušková (1990), Hušková (1998) Antoch and Hušková (1989), Hušková and Picek (2002), etc. The problem is the estimator $\hat{\sigma}_n^2(\psi)$ that will be discussed in the next subsection.

Large values of the above test statistics indicate that the null hypothesis is violated. We need at least approximations of the critical values with pre-chosen level α, i.e., approximations for the null distributions of the test statistics is needed. It can be obtained either through the limit null distribution or via bootstrap. The limit behavior under various sets of assumptions is studied in the following subsection.

11.2.2 Assumptions and Theoretical Results

The limit behavior of the above test statistics and their modifications have been studied under various assumptions on the score function ψ, regression vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots$ and the distribution of error terms $e_1, e_2, \ldots$. The considered assumptions on ψ are rather standard in the framework of robust statistics. We will show that the limit behavior of the test statistics sometimes depends on $\mathbf{x}_1, \mathbf{x}_2, \ldots$. Concerning assumptions on $e_1, e_2, \ldots$ we discuss here two main sets of assumptions. The first one assumes i.i.d. random errors which has been quite extensively studied in the literature. The other one admits dependence that is of recent interest.

At first, we formulate the assumptions on the score function ψ and the random error distribution function F:

(A.1) ψ are monotone (nondecreasing) functions, $\lambda_F(t) = -\int \psi(x-t)\,dF(x)$, $t \in \mathbb{R}$, $\lambda_F(0) = 0$, $\lambda'_F(0) > 0$, $\lambda'(t)$ exists in a neighborhood of 0 and is Lipschitz in a neighborhood of 0, e.g., for $|t| \le c_0$ with some $c_0 > 0$.

(A.2) $\int |\psi(t)|^{2+\Delta}\,dF(t) < \infty$ for some $\Delta > 0$ and

$$\int \big|\psi(x+t_2) - \psi(x+t_1)\big|^2\,dF\,j(x) \le C_1|t_2 - t_1|^{\kappa}, \quad |t_1|,\ |t_2| \le c_0$$

for some $1 \le \kappa \le 2$, $c_0, c_1 > 0$.

These assumptions are quite standard in robust statistics. For further information on the choice of ψ, see the classical works on robust methods, e.g., Jurečková and Sen (1996), Huber (1981).

Let us recall some of the most often considered ψ-functions. The classical choice $\psi(x) = x,\ x \in \mathbb{R}$, leads to the ordinary least squares (OLS) and L_2-residuals. A choice of $\psi(x) = \operatorname{sign} x,\ x \in \mathbb{R}$, leads to L_1-estimators and L_1-residuals, sometimes called LAD (least absolute deviation) procedures. Huber (1981) introduced $\psi(x) = xI\{|x| \le K\} + K \operatorname{sign} x I\{|x| > K\}, x \in \mathbb{R}$ for some $K > 0$, which is one of the most often used score functions, usually known as the Huber function.

Concerning the distribution of the error terms $e_1, \ldots, e_n$ it is assumed:

(B.1) $\{e_i\}_i$ is a sequence of i.i.d. random variables with common distribution function F.

We consider two basic types of assumptions on the regression vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Particularly, we assume that either $\mathbf{x}_i, \ldots, \mathbf{x}_n$ have "no trend" (see (C.1)–(C.3) below) or they have "a trend" (see (11.18) below). The former case is called *non-trending regression* while the later one is usually called *trending regression*.

The non-trending regression assumes:

(C.1) $x_{i1} = 1, i = 1, \ldots, n$, and $\sum_{i=1}^{n} x_{ij} = 0$, $j = 2, \ldots, p$.

(C.2) There exists a positive definite $p \times p$ matrix $\mathbf{C}$ such that for any sequence $\{l_n\}_n$ with the properties $l_n \le n$ and $\lim_{n\to\infty} l_n = \infty$, as $n \to \infty$,

$$\max_{1 \le k \le n - l_n} \left\| \frac{1}{l_n}(\mathbf{C}_{k+l_n} - \mathbf{C}_k) - \ \mathbf{C} \right\| = O\big((\log l_n)^{-1}\big)$$

uniformly in $1 \le k \le n - l_n$, where $\|.\|$ denotes the Euclidean norm.

(C.3) It holds, as $n \to \infty$,

$$\max_{1 \le k \le n} \left\{ \frac{1}{k} \sum_{i=1}^{k} \|\mathbf{x}_i\|^4 + \frac{1}{n-k} \sum_{i=k+1}^{n} \|\mathbf{x}_i\|^4 \right\} = O(1).$$

Next, the main assertions under H_0 for non-trending regression are stated.

Theorem 11.1 *Let* $(Y_1, \mathbf{x}_1), \ldots, (Y_n, \mathbf{x}_n)$ *follow the model* (11.1). *Moreover, let assumptions* (A.1)–(A.2), (B.1) *and* (C.1)–(C.3) *be satisfied. Let*

$$\hat{\sigma}_n^2(\psi) = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_{i,n}^2(\psi). \tag{11.14}$$

Then under H_0

$$\lim_{n\to\infty} P(T_{n,0}(\psi) \le t + 2\log\log n + p \log\log\log n - 2\log\big(2\Gamma(p/2)\big)$$
$$= \exp\big\{-2\exp\{-t/2\}\big\}, \quad t \in \mathbb{R}, \tag{11.15}$$

$$\lim_{n\to\infty} P\big(T_{n,1}(\psi) \le x\big) = P\left(\sup_{0<t<1} \sum_{j=1}^{p} B_j^2(t) \le x \right), \quad x \in \mathbb{R}, \tag{11.16}$$

$$\lim_{n\to\infty} P\big(T_{n,B}(\psi) \le x\big) = P\left(\int_0^1 \sum_{j=1}^{p} B_j^2(t)\, dt \le x \right), \quad x \in \mathbb{R}, \tag{11.17}$$

where $\{B_j(t); t \in (0,1)\}$, $j = 1, \ldots, p$, *are independent Brownian bridges and*

$$\Gamma(p) = \int_0^\infty t^{p-1} \exp\{-t\}\, dt.$$

Next, we turn to trending regression. To avoid quite complex formulations of the assumptions, we consider only polynomial regression:

$$\mathbf{x}_i = \big(1, i/n, \ldots, (i/n)^{p-1}\big)^T, \quad i = 1, \ldots, n. \tag{11.18}$$

Here is the related assertion.

Theorem 11.2 *Let* $(Y_1, \mathbf{x}_1), \dots, (Y_n, \mathbf{x}_n)$ *follow the model* (11.1). *Let* (A.1)–(A.2) *and* (B.1) *be satisfied and let* $\hat{\sigma}_n^2(\psi)$ *be defined in* (11.14). *Then under* H_0

$$\lim_{n\to\infty} P\left(T_{n,0}(\psi) \le t + 2\log\log n + p\log\log\log n - 2\log\left(\frac{2^{p/2}\Gamma(p/2)}{p}\right)\right)$$
$$= \exp\{-2\exp\{-t/2\}\}, t \quad \in \mathbb{R}, \qquad (11.19)$$

$$\lim_{n\to\infty} P\big(T_{n,1}(\psi) \le x\big) = P\Big(\sup_{0<t<1} \mathbf{S}^T(t)\mathbf{C}^{-1}(1)\mathbf{S}(t) \le x\Big), \quad x \in \mathbb{R}, \qquad (11.20)$$

$$\lim_{n\to\infty} P\big(T_{n,B}(\psi) \le x\big) = P\left(\int_0^1 \mathbf{S}^T(x)\mathbf{C}^{-1}(1)\mathbf{S}(x)\,dt \le x\right), \quad x \in \mathbb{R}, \quad (11.21)$$

where

$$\mathbf{S}(t) = \int_0^t \mathbf{h}(x)\,dW(x) - \mathbf{C}(t)\mathbf{C}(1)^{-1}\int_0^1 \mathbf{h}(x)\,dW(x),$$

$$\mathbf{C}(t) = \int_0^t \mathbf{h}(x)\mathbf{h}^T(x),\, dx, \quad t \in (0,1),$$

$$\mathbf{h}(x) = \big(1, x, \dots, x^{p-1}\big)^T, \quad x \in (0,1)$$

and $\{W_j(t); t \in (0,1)\}$, $j = 1, \dots, p$, *are independent Wiener processes.*

Remarks

- For $\psi(x) = x$, we get classical results known for L_2 procedures.
- The proof of Theorem 11.1 together with some related results for non-trending regression can be found, e.g., in Hušková (1996), while the proof of Theorem 11.2 can be found in Hušková and Picek (2002) and Aue et al. (2009). Hušková (2000, 2001) considered and studied a class of invariant robust tests. The assertions hold true even under weaker assumptions but formulation of the assumptions becomes quite complex.
- Under the assumptions of either theorems, the limit distributions depend neither on the choice of ψ and nor on the error distribution which means that the test statistics are asymptotically distribution free (under H_0).
- In case of non-trending regression (Theorem 11.1), the limit behaviors of the test statistics do not depend on the regression vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ while in case of polynomial trend this is not the case for $T_{n,1}(\psi)$ and $T_{n,B}(\psi)$ (Theorem 11.2).
- The test statistics $T_{n,1}(\psi)$ and $T_{n,B}(\psi)$ have the limit distribution as functionals of the Wiener processes. Therefore, approximations for the critical values can be obtained simulating the limit processes.
- The limit distribution of $T_{n,0}(\psi)$ belongs to the class of extreme values distributions. These results provide simple approximations for critical values, however the convergence is extremely slow.
- Bootstrap based on M-residuals gives good results both from theoretical results and simulations, see Hušková and Picek (2002) and Hušková and Picek (2004).

- $\hat{\sigma}_n^2(\psi)$ defined in (11.14) is an acceptable estimator of $\mathrm{var}\{\psi(e_1)\}$.
- It can be shown that if $\lim_{n\to\infty} \|\boldsymbol{\delta}_n\| \to 0$, $\lim_{n\to\infty} \|\boldsymbol{\delta}_n\|\sqrt{n} \to \infty$ and $k^* = \lfloor ns \rfloor$ for some $s \in (0, 1)$ then $\min(T_{n,1}(\psi), T_{n,B}(\psi)) \to^P \infty$. The same holds true for $T_{n,0}(\psi)$ if $\lim_{n\to\infty} \|\boldsymbol{\delta}_n\|\sqrt{n}(\log\log n)^{-1} = \infty$.
- Multiple changes. Marušiaková (2009) focused on linear models with trending regressors and independent errors. She generalized the L_2-tests proposed by Bai (1998) to M-type tests.

Next we give a general remark on the proofs of the limit behavior of our test statistics for ψ fulfilling (A.1), (A.2). There are three main steps for proving the assertions (11.16), (11.17), (11.20) and (11.21). To get the assertions on $T_{n,1}(\psi)$ and $T_{n,B}(\psi)$, we need to show:

- $$\max_{p\le k\le n-p} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{k} \mathbf{x}_i \hat{e}_{i,n}(\psi) - \left(\sum_{i=1}^{k} \mathbf{x}_i \psi(e_i) - \mathbf{C}_k \mathbf{C}_n^{-1} \sum_{j=1}^{n} \mathbf{x}_j \psi(e_i) \right) \right| = o_P(1),$$
- convergence (in $D(0, 1)$) of the process
$$\left\{ \mathbf{V}_n(t) = \mathbf{C}_n^{-1/2} \sum_{i=1}^{\lfloor n \rfloor} \mathbf{x}_i \psi(e_i), t \in (0, 1) \right\}$$
to the p-dimensional Wiener process with independent components,
- $$\hat{\sigma}_n^2(\psi) - \sigma^2(\psi) = o_P(1).$$

To show (11.15) and (11.19), we need stronger results, particularly, instead of the convergence of the process $\{\mathbf{V}_n(t), t \in (0, 1)\}$ a strong approximation with a certain rate is needed.

Now we turn to the situation when $\{Y_i, \mathbf{x}_i\}_i$ is a sequence of *weakly dependent random vectors*. We assume:

(D.1) For any $i \in \mathbb{Z}$, $\mathbf{x}_i = \mathbf{g}_1(\boldsymbol{\xi}_i, \boldsymbol{\xi}_{i-1}, \ldots)$, where $\mathbf{g}_1(\cdot)$ is a p-dimensional measurable function, $\{\boldsymbol{\xi}_i\}$ is a sequence of i.i.d. random vectors with dimension q_1, and $E\|\mathbf{x}_i\|^{2+\Delta} < \infty$ for some $\Delta > 0$.

(D.2) For any $i \in \mathbb{Z}$, $e_i = g_2(\boldsymbol{\zeta}_i, \boldsymbol{\zeta}_{i-1}, \ldots)$, where $g_2(\cdot)$ is a measurable function, $\{\boldsymbol{\zeta}_i\}$ is a sequence of i.i.d. random vectors with dimension q_2.

(D.3) The sequences $\{e_i\}$ and $\{\mathbf{x}_i\}$ are independent.

(D.4) For all $i \in \mathbb{Z}$,
$$\sum_{L=1}^{\infty} \|\mathbf{x}_i - \mathbf{x}_i^{(L)}\|_2 < \infty,$$
where
$$\mathbf{x}_i^{(L)} = h\big(\boldsymbol{\xi}_i, \boldsymbol{\xi}_{i-1}, \ldots \boldsymbol{\xi}_{i-L+1}, \boldsymbol{\xi}_{i-L}^{(L)}, \boldsymbol{\xi}_{i-L-1}^{(L)}, \ldots\big),$$

with

$$\boldsymbol{\xi}_{i-L}^{(L)}, \quad \boldsymbol{\xi}_{i-L-1}^{(L)}, \quad \dots$$

being i.i.d. with the same distribution as $\boldsymbol{\xi}_0$ and independent of $\{\boldsymbol{\xi}_i\}$.

(D.5) For $\psi(e_i)$, $i \in \mathbb{Z}$, the following is satisfied:

$$\sum_{L=1}^{\infty} \sup_{|a| \le a_0} \left\| \psi(e_i - a) - \psi\big(e_i^{(L)} - a\big) \right\|_2 < \infty$$

for some $a_0 > 0$, where

$$e_i^{(L)} = g_2\big(\boldsymbol{\zeta}_i, \boldsymbol{\zeta}_{i-1}, \dots \boldsymbol{\zeta}_{i-L+1}, \boldsymbol{\zeta}_{i-L}^{(L)}, \boldsymbol{\zeta}_{i-L-1}^{(L)}, \dots\big)$$

with

$$\boldsymbol{\zeta}_{i-L}^{(L)}, \quad \boldsymbol{\zeta}_{i-L-1}^{(L)}, \quad \dots$$

being i.i.d. with the same distribution as $\boldsymbol{\zeta}_0$ and independent of $\{\boldsymbol{\zeta}_i\}$.

Assumptions (D.1)–(D.2) correspond to Bernoulli shifts and (D.4)–(D.5) model dependence called $L_p - m$ approximability. This type of assumptions is coming from Hörmann and Kokoszka (2010) and Horváth and Kokoszka (2012) where the concept of $L_p - m$ approximability is introduced and studied. Discussion on relations to other type of dependence and as well as various particular cases are also presented there.

Here we consider the test statistic

$$T_{n,2}(\psi) = \sup_{0<t<1} \left\{ \frac{1}{n} \mathbf{S}_{\lfloor (n+1)t \rfloor}^T(\psi) \big(\hat{\boldsymbol{\Sigma}}_n(\psi)\big)^{-1} \mathbf{S}_{\lfloor (n+1)t \rfloor}(\psi) \right\}, \tag{11.22}$$

where $\mathbf{S}_k(\psi)$ is defined in (11.9),

$$\hat{\boldsymbol{\Sigma}}_n(\psi) - \boldsymbol{\Sigma}(\psi) = o_P(1) \tag{11.23}$$

with

$$\begin{aligned} \boldsymbol{\Sigma}(\psi) &= \lim_{n \to \infty} \operatorname{var} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{x}_i \psi(e_i) \right\} \\ &= E\big(\mathbf{x}_1 \mathbf{x}_1^T\big) E\psi^2(e_1) + 2 \sum_{v=1}^{\infty} E\big(\mathbf{x}_1 \mathbf{x}_v^T\big) E(\psi(e_1)\psi(e_{1+v}). \end{aligned} \tag{11.24}$$

Theorem 11.3 *Let* $\{Y_i, \mathbf{x}_i\}_i$ *follow model* (11.1)*. Let assumptions* (A.1)–(A.2), (D.1)–(D.5) *and* (11.23) *be satisfied. Moreover, let* $\boldsymbol{\Sigma}(\psi)$ *be positive definite. Then under* H_0

$$\lim_{n \to \infty} P\big(T_{n2}(\psi) \le x\big) = P\left(\sup_{0<t<1} \sum_{j=1}^{p} B_j^2(t) \le x \right), \quad x \in \mathbb{R}.$$

The variance matrix $\boldsymbol{\Sigma}(\psi)$ can be estimated as follows:

$$\hat{\boldsymbol{\Sigma}}_n(\psi) = \sum_{|k|<q} \big(1 - |k|/q\big) \hat{\boldsymbol{\Gamma}}_k,$$

where $q = q(n)$ and ω_q is a kernel function specified below, $\hat{\boldsymbol{\Gamma}}_k$ is the k-th lag sample covariance corresponding to $\boldsymbol{\Gamma}_k$, i.e.,

$$\hat{\boldsymbol{\Gamma}}_k = \begin{cases} \frac{1}{n} \sum_{i=1}^{n-k} \mathbf{x}_i \mathbf{x}_{i+k}^T \psi(\hat{e}_i)\psi(\hat{e}_{i+k}), & k \geq 0, \\ \hat{\boldsymbol{\Gamma}}_{-k}^T, & k < 0. \end{cases} \tag{11.25}$$

Remarks

- The proof of this theorem together with further results is found in Prášková and Chochola (2013). We survey here the basic results.
- Under mild conditions, the test is consistent.
- The properties of estimation of $\hat{\boldsymbol{\Sigma}}_n(\psi)$ are studied in Chochola et al. (2013).
- If $\{e_i\}_i$ is a sequence of i.i.d. random variables we can use

  $$\hat{\boldsymbol{\Sigma}}_n(\psi) = \frac{1}{n} \mathbf{C}_n \hat{\sigma}_n^2(\psi)$$

  with $\hat{\sigma}_n^2(\psi)$ defined in (11.14) and $T_{n2}(\psi)$ reduces to $T_{n1}(\psi)$.
- Similarly as in independent observations approximations for the critical values can be obtained either simulating asymptotic distribution or applying proper bootstrap (in this case circular block bootstrap works well).
- Another type of dependence was considered by Hušková and Marušiaková (2012).

11.3 Robust Estimators of a Change

This section concerns M-estimators of the change point in regression models with a single change as well as multiple changes. Results on consistency and limit distribution for local alternatives are discussed.

We start with regression models with a single change (11.1), nonzero $\boldsymbol{\delta}_n$ and

$$k^* = \lfloor ns \rfloor, \quad \text{for some } s \in (0, 1). \tag{11.26}$$

Motivated by the L_2-estimator $\hat{k}_n^*$ defined in (11.7) we introduce the M-estimator $\hat{k}_n^*(\psi)$ as follows:

$$\begin{aligned} \hat{k}_n^*(\psi) = \min\Big\{ & k;\, p < k < n-p,\, \mathbf{S}_k^T(\psi)\mathbf{C}_k^{-1}\mathbf{C}_n\big(\mathbf{C}_k^0\big)^{-1}\mathbf{S}_k(\psi) \\ & = \max_{p \leq j < n-p} \mathbf{S}_j^T(\psi)\mathbf{C}_j^{-1}\mathbf{C}_n\big(\mathbf{C}_j^0\big)^{-1}\mathbf{S}_j(\psi)\Big\}. \end{aligned} \tag{11.27}$$

Other estimators, e.g., the estimator related to $T_{n1}(\psi)$, can be also introduced, see Antoch and Hušková (1999, 2000).

Antoch and Hušková (2001) proved the following theorem.

Theorem 11.4 *Let* $\{Y_i, \mathbf{x}_i\}_i$ *follow model* (11.1). *Let assumptions* (A.1)–(A.2), (B.1)–(B.4) *and* (11.26) *be satisfied and let, moreover,*

$$\lim_{n\to\infty} \sqrt{\log n}\|\boldsymbol{\delta}_n\| \to 0, \qquad \lim_{n\to\infty} \|\boldsymbol{\delta}_n\|\sqrt{n/\log n} \to \infty. \tag{11.28}$$

Then

$$\kappa^2(\psi, F)\boldsymbol{\delta}_n^T \mathbf{C}\boldsymbol{\delta}_n\big(\hat{k}_n^*(\psi) - k^*\big) \to^D V,$$

where

$$\kappa^2(\psi, F) = \frac{(\lambda_F'(0))^2}{\int \psi^2(x)\,dF(x)},$$

$$V = \arg\max\big\{W_*(t) - |t|/2; t \in \mathbb{R}\big\}$$

with $\{W_*(t); t \in \mathbb{R}\}$ *being a two-sided Wiener process. The assertion remains true if* $\kappa^2(\psi, F)$ *is replaced by a consistent estimator.*

Remarks

- The explicit form of the distribution of V is known see, e.g., Csörgő and Horváth (1997).
- The assertion of Theorem 11.4 implies a consistency result:

$$n\|\boldsymbol{\delta}_n\|^2\big(\hat{k}_n^*(\psi) - k^*\big)/n = O_P(1).$$

- The limit distribution of $\hat{k}_n^*(\psi)$ depends on the error distribution F and the score function ψ through $\kappa^2(\psi, F)$. If F is known, then under some regularity conditions on F the score function ψ can be chosen in such a way that the asymptotic variance of $\hat{k}_n^*(\psi)$ is minima within the considered class of estimators.
- The related M-estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\beta} + \boldsymbol{\delta}_n$ can be determined from $Y_1, \dots, Y_{\hat{k}_n^*(\psi)}$ and $Y_{\hat{k}_n^*(\psi)+1}, \dots, Y_n$, respectively. They are $\sqrt{n}$-consistent.
- The alternative (11.28) is a local type. The limit distribution under fixed alternatives is more complex, for more information see Antoch and Hušková (1999) among others.
- The confidence interval can be constructed using the limit distribution and a consistent estimator of $\kappa^2(\psi, F)$. Another possibility is to use a bootstrap. For more details see Antoch and Hušková (1999) and Hušková and Kirch (2010).
- Other estimators and their properties for independent observations are studied, e.g., in Antoch and Hušková (1999, 2000, 2001).

Concerning related papers, under slightly different assumptions on $\mathbf{x}_1, \ldots, \mathbf{x}_n$ Bai (1995) studied limit properties of LAD-estimators of k^*, $\boldsymbol{\beta}$, $\boldsymbol{\beta} + \boldsymbol{\delta}_n$ defined as minimizers of

$$\sum_{i=1}^{k} \left| Y_i - \mathbf{x}_i^T \mathbf{b}_1 \right| + \sum_{i=k+1}^{n} \left| Y_i - \mathbf{x}_i^T \mathbf{b}_2 \right|$$

w.r.t. $k, \mathbf{b}_1, \mathbf{b}_2$. He showed that under local alternatives these estimators have the same type of the limit distribution as described in Theorem 11.4.

Fiteni (2002) considered a more general regression setup with dependent error terms (strong mixing) and studied limit properties of the estimators of k^*, $\boldsymbol{\beta}$, $\boldsymbol{\beta} + \boldsymbol{\delta}_n$ defined as minimizers of

$$\sum_{i=1}^{k} \rho\big(\big(Y_i - \mathbf{x}_i^T \mathbf{b}_1\big)/s_{1,k}\big) + \sum_{i=k+1}^{n} \rho\big(\big(Y_i - \mathbf{x}_i^T \mathbf{b}_2\big)/s_{k+1,n}\big)$$

w.r.t. $k, \mathbf{b}_1, \mathbf{b}_2$. Here ρ is a convex loss function satisfying (A.1)–(A.2), $s_{1,k}$ and $s_{k+1,n}$ are scale estimators based on $Y_1, \ldots, Y_k$ and $Y_{k+1}, \ldots, Y_n$, respectively. These estimators of the change point generally differ from $\hat{k}_n^*(\psi)$ but the asymptotic behavior for i.i.d. error terms and for local alternatives coincides. The estimators proposed by Fiteni (2002) are scale invariant which is not generally true for $\hat{k}_n^*(\psi)$.

There are also results for multiple changes. Toward this we consider multiple change regression model:

$$Y_i = \sum_{j=1}^{q+1} \sum_{i=k_{j-1}^*}^{k_j^*-1} \mathbf{x}_i^T \boldsymbol{\beta}_j + e_i, \quad i = 1 \ldots, n, \tag{11.29}$$

where q is total number of changes, $1 = k_0^* < k_1^* < \cdots < k_{q+1}^* = n + 1$, $\boldsymbol{\beta}_j j = 1, \ldots, q$, are p-dimensional unknown regression parameters such that $\boldsymbol{\beta}_j \neq \boldsymbol{\beta}_{j+1}, j = 1, \ldots, q, \mathbf{x}_i = (x_{i,1}, \ldots, x_{ip})^T, x_{i,1} = 1, i = 1, \ldots, n$ are known regression vectors and $e_1, \ldots, e_n$ are again random errors satisfying certain regularity conditions. Bai (1998) introduced LAD-type estimators for this situation. He defined the estimators of $k_1^* < \cdots < k_q^*$, $\boldsymbol{\beta}_j j = 1, \ldots, q$ as minimizers of

$$\sum_{j=1}^{q+1} \sum_{i=k_{j-1}+1}^{k_j} \left| Y_i - \mathbf{x}_i^T \mathbf{b}_j \right|$$

w.r.t. $k_1, \ldots, k_q, \mathbf{b}_1, \ldots, \mathbf{b}_{q+1}$ under side condition $k_j^* - k_{j-1}^* \geq cn^{3/4}$ for some $c > 0$. It is shown that the proposed estimators are consistent under mild assumptions. Additionally under fixed alternatives the limit distribution of the estimators is derived and further results are discussed.

In practice an increasing interest is payed to estimation of the number of changes, i.e., when also q is unknown in (11.29). Bai (1998) proposed LAD type estimator with penalty term. Particularly, he proposed to estimate q as the minimizer of

$$\min_{*} \frac{1}{n} \sum_{j=1}^{q+1} \sum_{i=k_{j-1}}^{k_j - 1} \left| Y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{j,q} \right| + \frac{1}{2}(q+1)(p+1)\log n,$$

where $\min_*$ is the minimum over the set $\{1 = k_0 < k_1 < \cdots < k_q < k_{q+1} = n, k_j - k_{j-1} \geq n^{\gamma}\}$ with properly chosen $0 < \gamma \leq 1$ and $\hat{\boldsymbol{\beta}}_{j,q}$ is LAD estimator of $\boldsymbol{\beta}_j$ based on $Y_{k_{j-1}}, \ldots Y_{k_j - 1}$, $j = 1, \ldots, q-1$. Such estimators of the total number of changes are consistent. Ciuperca (2011a) and Ciuperca (2011b) considered and studied LAD type estimators in nonlinear regression with multiple changes both with known and unknown number of changes. To get an estimator of the number of changes for penalization a modified Schwarz criterion is used. Ciuperca (2009) deals with M-estimators in nonlinear regression with multiple changes. In all mentioned papers on multiple changes the error terms are supposed to be i.i.d. with some additional properties.

11.4 Miscellaneous

11.4.1 Sequential Robust Procedures

We shortly mention a class of sequential robust procedures for detection of linear regression model

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta}_i + e_i, \quad 1 \leq i < \infty, \tag{11.30}$$

with possible changes in the p-dimensional regression parameters $\boldsymbol{\beta}_i$, $1 \leq i < \infty$, $\{\mathbf{x}_i\}_i$ are known p-dimensional regression vectors and $\{e\}_i$ is a sequence of the error terms satisfying certain regularity assumptions. It is assumed that a training sample of size m with no instabilities is available, i.e., $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_m$ and the observations $Y_1, \ldots, Y_m$ represent the training period (historical data). These observations are used for calibration of the model and used in monitoring afterwards. It is assumed that the data are arriving sequentially.

Detection of a change in the regression model is formulated as a sequential hypothesis testing problem, where the null hypothesis H_0^S corresponds to the model without any change, i.e.,

$$H_0^S: \quad \boldsymbol{\beta}_i = \boldsymbol{\beta}_0, \quad 1 \leq i < \infty,$$

and the alternative hypothesis H_A reflects that the model changes at some unknown time-point, that is

$$\begin{aligned} H_A^S: \quad & \text{there exists } k^* \geq 1 \text{ such that } \boldsymbol{\beta}_i = \boldsymbol{\beta}_0,\, 1 \leq i < m + k^*, \text{ but} \\ & \boldsymbol{\beta}_i = \boldsymbol{\beta}_0 + \boldsymbol{\delta}_m,\ m + k^* \leq i < \infty, \quad \boldsymbol{\delta} \neq \mathbf{0}, \end{aligned}$$

where $\boldsymbol{\beta}_0, \boldsymbol{\delta} = \boldsymbol{\delta}_m$ and $k^* = k_m^*$ are unknown parameters.

For the detection of changes in the above model, Chu et al. (1996), Berkes et al. (2004), Horváth et al. (2004), Zeileis et al. (2005), Leisch et al. (2000), Aue et al. (2006), etc. studied various procedures based on L_2-estimators and L_2-residuals.

Now, we introduce robust sequential analogs of some test developed in the above mentioned papers. Our procedures will be based on the M-estimators $\hat{\boldsymbol{\beta}}_m(\psi)$ based on $Y_1, \ldots, Y_m$ and M-residuals $\hat{e}_{1,m}(\psi), i \geq 1$ defined in (11.8). Particularly, we use the quadratic forms of the partial sums of weighted M-residuals

$$Q(k,m;\psi) = \left(\frac{1}{\sqrt{m}} \sum_{i=m+1}^{m+k} \mathbf{x}_i \hat{e}_{i,m}(\psi)\right)^T \left(\hat{\boldsymbol{\Sigma}}_m(\psi)\right)^{-1} \left(\frac{1}{\sqrt{m}} \sum_{i=m+1}^{m+k} \mathbf{x}_i \hat{e}_{i,m}(\psi)\right), \quad k \geq 1,$$

where $\hat{\boldsymbol{\Sigma}}_m(\psi)$ is an estimator of the asymptotic variance of

$$\boldsymbol{\Sigma}(\psi) = \lim_{m\to\infty} \operatorname{var}\left\{\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \mathbf{x}_i \psi(e_i)\right\}.$$

The null hypothesis is rejected as soon as for some k

$$Q(k,m;\psi)/q_\gamma(k/m) \geq c,$$

for an appropriately chosen $c = c_\gamma(\alpha)$, where $q_\gamma(t), t \in (0,\infty)$ is a suitable boundary (weight) function. In this case, we stop the procedure and confirm a change, otherwise we continue monitoring. The associated stopping rule is given by

$$\tau_m = \tau_m(\gamma) = \inf\left\{1 \leq k < Tm : Q(k,m,\psi)/q_\gamma(k/m) \geq c\right\}$$

for $T > 0$. An approximation for the tuning constant c can be obtained from the limit behavior of $\max_{1\leq k\leq mT} Q(k,m,\psi)/q_\gamma(k/m)$ as $m \to \infty$. It can be shown that under (A.1)–(A.2), (D.1)–(D.5) and (11.23) with n replaced by m and under H_0^S

$$\begin{aligned} \lim_{m\to\infty} & P\left(\max_{1\leq k\leq mT} Q(k,m,\psi)/q_\gamma(k/m) \leq x\right) \\ &= P\left(\sup_{0<t<T/(T+1)} \left(\sum_{j=1}^{p} W_j^2(t)/t^{2\gamma}\right) \leq x\right), \quad \forall x, \end{aligned}$$

for $T > 0$, where $W_j(t)$, $j = 1, \dots, p, t \in (0, 1)$ are independent Brownian motions and

$$q_\gamma(t) = (1+t)^2\big(t/(t+1)\big)^{2\gamma}, \quad t \in \mathbb{R}.$$

The proof together with further limit properties, simulations and applications are in Koubková (2006) for i.i.d. errors and in Chochola et al. (2013) for weakly dependent observations.

11.4.2 Rank Based Procedures

So far we have solely focused on M-procedures for detection of changes in regression models. Among robust methods belong also rank based procedures as well as procedures based on U-statistics. Both types were developed and studied mostly for independent observations. Probably, rank based methods were developed sooner than M-type ones.

We mentioned here only the very simple situation when $X_1, \dots, X_n$ are independent random variables, X_i has a continuous distribution function and we are interested in testing:

$$H_0: \quad F_1 = \dots = F_n$$

against the alternative

$$H_A: \quad \text{there is } k^* < n \text{ such } F_1 = \dots = F_{k^*} \neq F_{k^*+1} = \dots = F_n,$$

where k^* is unknown. As test statistics one can use functionals of the simple linear rank statistics

$$S_k^R = \sum_{i=1}^{k}\big(a(R_i) - \bar{a}_n\big), \quad k = 1, \dots, n,$$

where $R_1, \dots, R_n$ are ranks of $X_1, \dots, X_n$, $a(1), \dots, a(n)$ are scores and $\bar{a}_n = \frac{1}{n}\sum_{i=1}^{n} a(i)$.

Similarly to the M-type procedures the following types of test procedures were developed and studied:

$$T_{n,0}^R = \max_{1 \le k < n}\left\{\sqrt{\frac{n}{k(n-k)}}\frac{1}{\sigma_{n,a}}\left|S_k^R\right|\right\},$$

$$T_{n,1}^R = \max_{1 \le k < n}\left\{\frac{1}{\sqrt{n}}\frac{1}{\sigma_{n,a}}\left|S_k^R\right|\right\},$$

where

$$\sigma_{n,a}^2 = \frac{1}{n}\sum_{i=1}^{n}\big(a(i) - \bar{a}_n\big)^2.$$

Large values of the test statistics indicate that the null hypothesis is violated. These test statistics are distribution free under the null hypothesis and therefore approximations for the desired critical values can be simulated. Concerning their asymptotic properties Hušková (1997a) and Hušková (1997b) proved that if $X_1, \dots, X_n$ are i.i.d. with common continuous distribution function and additionally

$$\liminf_{n\to\infty} \sigma_{n,a}^2 > 0, \qquad \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \left|a(i) - \bar{a}_n\right|^{2+\delta} < \infty$$

for some positive δ, then

$$\lim_{n\to\infty} P\left(\sqrt{2\log\log n}\, T_{n,0}^R \le t + 2\log\log n + \frac{1}{2}\log\log\log n - \frac{1}{2}\log \pi\right)$$

$$= \exp\left\{-2e^{-t}\right\},$$

$$\lim_{n\to\infty} P\left(T_{n,1}^R \le t\right) = P\left(\sup_{0<s<1} \left|B(s)\right| \le t\right),$$

where $t \in \mathbb{R}$ and $\{B(s), s \in (0,1)\}$ is a Brownian bridge on $(0,1)$. The estimator of the change point k^* related to $T_{n,0}^R$ is defined as k maximizing $\sqrt{\frac{n}{k(n-k)}}\frac{1}{\sigma_{n,a}}|S_k^R|$. It is introduced and studied in Gombay and Hušková (1998).

The results for simple models were extended to some regression models with independent errors. P.K. Sen contributed remarkably to the rank based procedures for detection changes in eighties, see survey papers Sen (1991) and Hušková and Sen (1989). Antoch et al. (2008) developed and studied data driven rank based procedures.

U-statistics procedures were introduced and studied, e.g., by Horváth and Gombay (1995), Gombay (2000a, 2000b, 2001, 2004) and Horváth and Hušková (2005).

11.5 Conclusions

The area of change-point problem is fast developing since there are many situations in the real world with various changes. Presently, the main focus is on changes in regression models and time series covering not only changes in regression parameters but also in structural dependencies. High interest is in modeling changes in multivariate and functional data. Most of these procedures are related to L_2 procedures. Due to their sensitivity w.r.t. outliers or to heavy tailed error distributions it is deserved to develop and to study robust modifications. L_2 procedures can wrongly classify outliers as a change in the model.

Big interest is also in detection of multiple changes, particularly, in estimation of the number of change points and their locations based on robust approach.

Statistical procedures for detection of changes are computationally complex even for the L_2 type procedures, however robust procedures are computationally still more complex. There is a need of efficient algorithms.

Acknowledgements We gratefully acknowledge GAČR P201/12/1277 for support.

References

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, *61*, 821–856.

Aue, A., Horváth, L., Hušková, M., & Kokoszka, P. (2006). Change-point monitoring in linear models. *Econometrics Journal*, *9*, 373–403.

Aue, A., Horváth, L., & Hušková, M. (2009). Extreme value theory for stochastic integrals of Legendre polynomials. *Journal of Multivariate Analysis*, *100*, 1029–1043.

Antoch, J., & Hušková, M. (1989). Some M-tests for detection of a change in linear models. In P. Mandl & M. Hušková (Eds.), *Proceedings of the fourth Prague symposium on asymptotic statistics* (pp. 123–136).

Antoch, J., & Hušková, M. (1999). Estimators of changes. In *Statistics: a series of textbooks and monographs: Vol. 158. Asymptotics, nonparametrics and time series* (pp. 533–577). New York: Dekker.

Antoch, J., & Hušková, M. (2000). Bayesian like R- and M-estimators of change points. *Discussiones Mathematicae*, *20*, 114–134.

Antoch, J., Hušková, M., & Jarušková, D. (2000). Change point analysis. In *Lecture notes from the 5th IASC Summer School* (pp. 1–76).

Antoch, J., & Hušková, M. (2001). M-estimators of structural changes in regression models. *Tatra Mountains Mathematical Publications*, *22*, 197–208.

Antoch, J., Hušková, M., Janic, A., & Ledwina, T. (2008). Data driven rank test for detection of changes. *Metrika*, *68*, 1–15.

Bai, J. (1995). Least absolute deviation estimation of a shift. *Econometric Theory*, *11*, 403–436.

Bai, J. (1998). Estimation of multiple-regime regressions with least absolutes deviation. *Journal of Statistical Planning and Inference*, *74*, 103–134.

Bai, J., & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, *66*, 47–78.

Basseville, M., & Nikiforov, I. V. (1993). *Detection of abrupt changes: theory and applications*. New York: Prentice Hall.

Berkes, I., Gombay, E., Horváth, L., & Kokoszka, P. (2004). Sequential change-point detection in GARCH(p, q) models. *Econometric Theory*, *20*, 1140–1167.

Brodsky, B. S., & Darkhovsky, D. E. (1993). *Nonparametric methods in change-point problems*. Dordrecht: Kluwer Academic.

Carlstein, E., Müller, H.-G., & Siegmund, D. (1994) In *IMS lecture notes monograph series: Vol. 23. Change-point problems*. Hayward: Institute of Mathematical Statistics.

Chen, J., & Gupta, A. K. (2000). *Parametric statistical change point analysis*. Boston: Birkhäuser Boston

Chochola, O., Hušková, M., Prášková, Z., & Steinebach, J. (2013). Robust monitoring of CAPM portfolio betas. *Journal of Multivariate Analysis*, *115*, 374–395.

Chu, C.-S. J., Stinchcombe, M., & White, H. (1996). Monitoring structural change. *Econometrica*, *64*, 1045–1065.

Ciuperca, G. (2009). The M-estimation in multiple-phase random nonlinear model. *Statistics & Probability Letters*, *75*, 573–580.

Ciuperca, G. (2011a). Estimating nonlinear regression with and without change points by the LAD. *Annals of the Institute of Statistical Mathematics*, *63*, 717–743.

Ciuperca, G. (2011b). Penalized least absolute deviations estimation for nonlinear model with change-points. *Statistical Papers*, *52*, 371–390.

Csörgő, M., & Horváth, L. (1997). *Limit theorems in change-point analysis*. Chichester: Wiley.

Fiteni, I. (2002). Robust estimation of structural break points. *Econometric Theory*, *18*, 349–386.

Gombay, E. (2000a). Comparison of U-statistics in the change-point problem and in sequential change detection. *Periodica Mathematica Hungarica*, *41*, 157–166.

Gombay, E. (2000b). U-Statistics for sequential change-detection. *Metrika*, *54*, 133–145.

Gombay, E. (2001). U-Statistics for change under alternatives. *Journal of Multivariate Analysis*, *78*, 139–158.

Gombay, E. (2004). U-statistics in sequential tests and change detection. *Sequential Analysis*, *23*, 254–274.

Gombay, E., & Hušková, M. (1998). Rank based estimators of the change point. *Journal of Statistical Planning and Inference*, *67*, 137–154.

Horváth, L., & Gombay, E. (1995). An application of U-statistics to change-point analysis. *Acta Scientiarum Mathematicarum*, *60*, 345–357.

Horváth, L., & Hušková, M. (2005). Testing for changes using permutations of U-statistics. *Journal of Statistical Planning and Inference*, *128*, 351–371.

Horváth, L., Hušková, M., Kokoszka, P., & Steinebach, J. (2004). Monitoring changes in linear models. *Journal of Statistical Planning and Inference*, *126*, 225–251.

Horváth, L., & Kokoszka, P. (2012). *Inference for functional data with applications*. Berlin: Springer.

Hörmann, S., & Kokoszka, P. (2010). Weakly dependent functional data. *The Annals of Statistics*, *38*, 1845–1884.

Huber, P. (1981). *Robust statistics*. Chichester: Wiley.

Hušková, M. (1990). Some asymptotic results for robust procedures for testing the constancy of regression models over time. *Kybernetika*, *26*, 392–403.

Hušková, M. (1996). Tests and estimators for change point problem based on M-statistics. *Statistics & Decisions*, *14*, 115–136.

Hušková, M. (1997a). Limit theorems for rank statistics. *Statistics & Probability Letters*, *32*, 45–55.

Hušková, M. (1997b). Multivariate rank statistics processes and change point analysis. In S. E. Ahmed, M. Ahsanullah, & B. K. Sinha (Eds.), *Applied statistical sciences III* (pp. 83–96). New York: Nova Science Publishers.

Hušková, M. (1998). L_1- test procedures for detection of change. In Y. Dodge (Ed.), *Lecture notes of IMS: Vol. 31. L_1-Statistical procedures and related topics* (pp. 57–70).

Hušková, M. (2000). Some invariant tests for detection of structural changes. *Kybernetika*, *36*, 401–414.

Hušková, M. (2001). Some invariant test procedures for detection of structural changes; behavior under alternatives. *Kybernetika*, *37*, 669–684.

Hušková, M., & Sen, P. K. (1989). Nonparametric tests for shift and change in regression at an unknown time point. In P. Hackl (Ed.), *Statistical analysis and forecasting of economic structural change* (pp. 73–87). Berlin: Springer.

Hušková, M., & Picek, J. (2002). M-tests for detection of structural changes in regression. In Y. Dodge (Ed.), *Statistical data analysis based on the L_1-norm and related methods* (pp. 213–229). Basel: Birhäuser.

Hušková, M., & Picek, J. (2004). Some remarks on permutation type tests in linear models, in regression models. *Discussiones Mathematicae. Probability and Statistics*, *24*, 151–181.

Hušková, M., & Kirch, K. (2010). A note on studentized confidence intervals for the change-point. *Computational Statistics*, *25*, 269–289.

Hušková, M., & Marušiaková, M. (2012). M-Procedures for detection of changes for dependent observations. *Communications in Statistics. Simulation and Computation*, *41*, 1032–1050.

Jurečková, J., & Sen, P. K. (1996). *Robust statistical procedures: asymptotic and interrelations*. New York: Wiley.

Koubková, A. (2006). *Sequential change-point analysis*. Ph.D. Thesis, Charles University in Prague.

Leisch, F., Hornik, K., & Kuan, C.-M. (2000). Monitoring structural changes with the generalized fluctuation test. *Econometric Theory*, *16*, 835–854.

Marušiaková (2009). *Tests for multiple changes in linear regression models*. Ph.D. Thesis, Charles University in Prague.

Perron, P. (2006). Structural change. In S. Durlauf & L. Blume (Eds.), *The new Palgrave dictionary of time series* (2nd ed.). Palgrave: Macmillan.

Prášková, Z., & Chochola, O. (2013). M-procedures for detection of a change under weak dependence. Submitted.

Sen, P. K. (1991). Nonparametric methods sequential analysis. In B. K. Ghosh & P. K. Sen (Eds.), *Handbook of sequential analysis* (pp. 331–362). New York: Dekker.

Zeileis, A., Leisch, F., Kleiber, C., & Hornik, K. (2005). Monitoring structural change in dynamic econometric models. *Journal of Applied Econometrics*, *20*, 99–121.

Chapter 12
Robust Signal Extraction from Time Series in Real Time

Matthias Borowski, Roland Fried, and Michael Imhoff

12.1 Introduction

Data streams are measured in many fields. For example, in intensive care the patient status is continuously assessed by online-monitoring systems which measure multiple vital parameters such as blood pressure and heart rate with high sampling frequencies of more than one observation per second. A high sampling frequency typically leads to noisy and outlier-contaminated time series. Moreover, the resulting time series are often not stationary but exhibit changing trends and level shifts as well as a changing variability of the noise, cf. the grey time series in Fig. 12.1. The underlying time-varying trend constitutes a signal which carries relevant information. An obvious but challenging goal is then to separate this signal from noise and outliers in real time, see Fig. 12.1. The signal contains level shifts and trends and is filtered by the so-called *Slope Comparing Adaptive Repeated Median* (SCARM) filter, explained in Sect. 12.4.

All signal extraction procedures (also called *filters*) explained here are based on the general assumption that the data y_t come from of a signal which is overlaid with additive noise and outliers:

$$Y_t = \mu_t + \varepsilon_t + \eta_t, \quad t \in \mathbb{Z}. \tag{12.1}$$

The signal μ_t is assumed to be smooth, but it may exhibit changing trends and occasional level shifts. The noise is represented by the independent error variables

M. Borowski (✉) · R. Fried
Faculty of Statistics, TU Dortmund University, 44221 Dortmund, Germany
e-mail: borowski@statistik.tu-dortmund.de

R. Fried
e-mail: fried@statistik.tu-dortmund.de

M. Imhoff
Medizinische Fakultät, Ruhr-Universität Bochum, 44780 Bochum, Germany
e-mail: mike@imhoff.de

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_12, © Springer-Verlag Berlin Heidelberg 2013

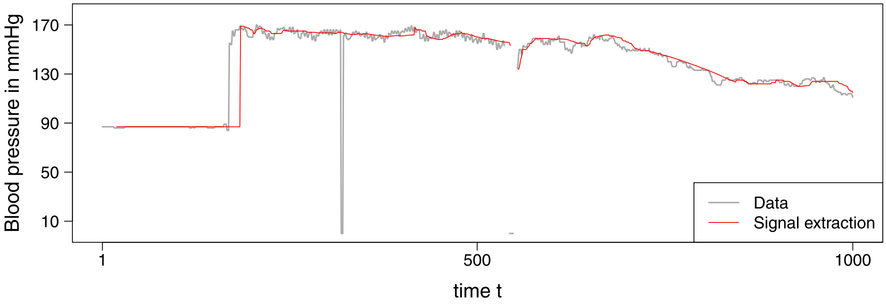


Fig. 12.1 Blood pressure in mmHg of a patient on an intensive care unit (*grey*) and signal extracted by the SCARM filter (*red*)

ε_t which have a symmetric distribution with expectation zero and time-dependent variance σ_t^2. The outlier process η_t generates impulsive, spiky noise but is equal to zero most of the time.

Since the signal changes over time, we use rather small moving time windows consisting of the latest n observations to approximate it in real time. A signal estimator which is applied to the window sample then has to meet several demands. It must be robust against single outliers and patches of outliers. In addition, it should yield a good efficiency in the sense of a small bias and variability. Signal changes like changing trends or level shifts should be traced accurately. Moreover, the filter is only applicable if the required computing time is shorter than the sampling frequency, which is why the method must be computationally fast.

In Sect. 12.2, several methods for real time signal extraction are explained which are based on robust estimators of location. In Sect. 12.3, we present robust regression-based estimators. The *Repeated Median* (RM, Siegel 1982) has turned out to deliver a good overall performance compared to other robust regression techniques. The window width n has a large influence on the resulting signal estimation, but it is difficult to choose an optimal window width for the RM in a real time application. We explain existing RM-based filters with automatic and time-dependent window width selection in Sect. 12.4. Finally, we present existing RM-based filters for online signal extraction from multivariate time series in Sect. 12.5. A summary and an outlook are provided in Sect. 12.6.

12.2 Location-Based Signal Extraction

Filtering procedures applying a location estimator to a moving window generally involve the assumption of a locally constant signal. The *running median* (see Tukey 1977, p. 210f) estimates the signal μ_t in a time window $\{t-w,\dots,t,\dots,t+w\}$ of width $n = 2w + 1$ by the median

$$\hat{y}_t^{\text{med}} = \text{med}\{y_{t-w},\dots,y_t,\dots,y_{t+w}\}$$

of the window observations, where the median is defined as the mean of the $(n/2)$-th and $((n/2)+1)$-th order statistic if n is even. The running median is highly robust against outliers and offers a *finite-sample replacement breakdown point* (fsbp, Donoho and Huber 1983) of $\lceil n/2 \rceil / n$ (where $\lceil c \rceil$ denotes the smallest integer not smaller than c). That is, $\lceil n/2 \rceil$ of the n observations in a sample of size n need to be replaced to achieve that the median estimator 'breaks down', meaning that it delivers an arbitrarily extreme value, roughly speaking. The running median traces level shifts well but deteriorates in trend periods, as trends are reproduced by steps. For median concepts for uni- and multivariate data, see the contributions by Oja (Chap. 1) and Rousseeuw and Hubert (Chap. 4). Details regarding the concept of breakdown points can be found in the contribution by Müller, Chap. 5.

The *Modified Trimmed Mean* (MTM, Lee and Kassam 1985) combines the robustness and shift preservation of the running median with the smoothness and the good efficiency of the non-robust running mean under Gaussian noise. The MTM first calculates the local median $\hat{y}_t^{\text{med}}$ of all observations in the time window $\{t-w, \dots, t, \dots, t+w\}$ and then trims, i.e., discards, the observations which deviate by more than a specified multiple of a robust estimate of the error scale. This scale estimator can be taken as the *median absolute deviation about the median* (MAD):

$$\hat{\sigma}_t^{\text{MAD}} = c_n \cdot \underset{i \in \{-w,\dots,w\}}{\text{med}} \left| y_{t+i} - \hat{y}_t^{\text{med}} \right|,$$

where the factor c_n depends on the sample size n and ensures unbiasedness at a specific error distribution. The arithmetic mean of the trimmed sample is then the MTM estimate of the signal at the central window time point t:

$$\hat{\mu}_t^{\text{MTM}} = \frac{1}{|I_t|} \sum_{i \in I_t} y_{t+i},$$

$$I_t = \left\{ i \in \{-w, \dots, w\} : \left| y_{t+i} - \hat{y}_t^{\text{med}} \right| \le d_t \right\}.$$

Lee and Kassam (1985) suggest using $d_t = 2\hat{\sigma}_t^{\text{MAD}}$ to achieve reasonable robustness and efficiency for Gaussian noise. Note that for $d_t = 0$ the MTM equals the running median, and for $d_t = \infty$ it equals the running mean.

Double window modified trimmed means (DWMTM, Lee and Kassam 1985) use the median and MAD in a shorter inner window $\{t-v, \dots, t, \dots, t+v\}$ with $v < w$ as initial estimator. The whole sample of size $n = 2w + 1$ is then trimmed based on these estimates, before the arithmetic mean is calculated on the trimmed sample. Using a short inner window improves shift preservation but reduces noise attenuation.

Linear median hybrid filters (Heinonen and Neuvo 1987, 1988) combine linear and median filters by applying linear subfilters to subsamples of the data and then taking the median of the subfilter outcomes as final filter output. Such a filter is called *Linear median hybrid filter with finite impulse response*, briefly FMH filter, if all linear subfilters $\phi_1, \dots, \phi_M$ give non-zero weight to a finite number of observations:

$$\hat{\mu}_t^{\mathrm{FMH}} = \mathrm{med}\big\{\phi_1(t), \ldots, \phi_M(t)\big\}.$$

The subfilters reduce the computation time as compared to the running median with the same window width. Moreover, the subfilters can be chosen such that polynomial trends of different degrees p are traced well, making the class of FMH filters very flexible. Several variations of FMH filters have been proposed, see Heinonen and Neuvo (1987, 1988), Astola et al. (1989), Wichman et al. (1990), for instance. To keep things short, we will only explain *Simple FMH* (SFMH) and *Predictive FMH* (PFMH) filters. SFMH filters assume the signal to be locally constant ($p = 0$) and use $M = 3$ subfilters, namely two arithmetic means and the central window observation y_t:

$$\phi_1(t) = \frac{1}{w}\sum_{i=1}^{w} y_{t-i}, \quad \phi_2(t) = y_t, \quad \phi_3(t) = \frac{1}{w}\sum_{i=1}^{w} y_{t+i}.$$

Taking the central observation as the subfilter $\phi_2(t)$ reduces the bias of the SFMH at level shifts as compared to the running median (Astola et al. 1989).

PFMH filters use weighted means instead of simple arithmetic means:

$$\hat{\mu}_t^{\mathrm{PFMH}} = \mathrm{med}\big\{\phi_A(t), y_t, \phi_B(t)\big\},$$

$$\phi_A(t) = \sum_{i=1}^{w} h_i y_{t-i}, \quad \phi_B(t) = \sum_{i=1}^{w} h_i y_{t+i}.$$

The weights h_i can be chosen in such a way that the *mean squared error* (MSE) is minimized, depending on the degree p of an underlying polynomial trend. For a linear trend ($p = 1$) and Gaussian noise, the weights $h_i = (4w - 6i + 2)/(w^2 - w)$, $i = 1, \ldots, w$, lead to signal estimates with minimal MSE (Heinonen and Neuvo 1988). Although the PFMH is based on location estimators, it can be adopted to linear signals with nonzero trends. However, if the signal is assumed to be locally linear, it is generally a better approach to achieve signal extraction by fitting regression lines.

12.3 Regression-Based Signal Extraction

Davies et al. (2004) propose the application of robust regression estimators in a moving time window $\{t - w, \ldots, t, \ldots, t + w\}$ of odd width $n = 2w + 1$. The estimate of the signal μ_t is then the level of the regression line at the central design point t. In a real time application, the signal would obviously be estimated with a delay of w time points. This approach can therefore be called *delayed* signal extraction. Since a time delay is often not desirable in an online-application, Gather et al. (2006) take the level of the regression line at the rightmost window point $t + w$ as signal estimate for that time point, so that the signal is estimated without time delay. This approach can thus be called *online* signal extraction. The online approach also allows for even window widths, since the signal is estimated (at the rightmost window point t) in the time window $\{t-n+1,\dots,t\}$ (Schettlinger et al. 2010).

Both the delayed and online approach assume that the signal can be approximated well by fitting regression lines to the most recent observations. This local linearity assumption can be expressed as follows:

$$\mu_{t+i} \approx \mu_t + \beta_t \cdot i, \quad i \in I, \tag{12.2}$$

where μ_{t+i} is the signal at time $t+i$ and β_t the slope in the time window. For the delayed approach, $I = \{-w,\dots,w\}$, while for the online approach, $I = \{-n+1,\dots,0\}$. In both cases, we aim at estimating the signal μ_t and the slope β_t by means of regression techniques. In the following, we explain several robust regression estimators. Given a sample $\mathbf{y}_t = (y_{t+i})_{i\in I}$ and a regression estimate $(\hat{\mu}_t, \hat{\beta}_t)$, we denote the residuals of the regression fit by

$$\mathfrak{r}_{t,i} = y_{t+i} - (\hat{\mu}_t + i\hat{\beta}_t), \quad i \in I.$$

L_1 regression (see, e.g., Rousseeuw and Leroy 1987, p. 10) minimizes the sum of the absolute values of the residuals:

$$\left(\hat{\mu}_t^{L_1}, \hat{\beta}_t^{L_1}\right) = \operatorname{argmin}\left\{(\hat{\mu}_t, \hat{\beta}_t) : \sum_{i\in I} |\mathfrak{r}_{t,i}|\right\}.$$

Calculation of L_1 regression is fast (e.g., Sposito 1990), but it may deliver multiple solutions. For large samples with design points on a lattice, the fsbp of the L_1 estimator is approximately $1 - 1/\sqrt{2} \approx 0.293$ (Davies et al. 2004).

Least Median of Squares regression (LMS, Hampel 1975; Rousseeuw 1984) offers larger robustness than L_1 regression. It has a fsbp of $\lfloor n/2 \rfloor / n \approx 50\ \%$ for samples of size n, which is the maximal breakdown point for regression equivariant estimators (Davies and Gather 2005). The LMS minimizes the median of the squared residuals:

$$\left(\hat{\mu}_t^{\mathrm{LMS}}, \hat{\beta}_t^{\mathrm{LMS}}\right) = \operatorname{argmin}\left\{(\hat{\mu}_t, \hat{\beta}_t) : \operatorname*{med}_{i\in I}\left\{\mathfrak{r}_{t,i}^2\right\}\right\}.$$

The LMS shows good resistance against many large outliers, since even almost 50 % outliers of any value do not cause a large bias. This is the reason why the LMS is almost unbiased at level shifts in the delayed setting. However, the LMS is computationally expensive and the LMS filter output is very wiggly because of its high variability.

The *Deepest Regression* (DR, Rousseeuw and Hubert 1999) estimator is given by

$$\left(\hat{\mu}_t^{\mathrm{DR}}, \hat{\beta}_t^{\mathrm{DR}}\right) = \operatorname{argmax}\left\{(\hat{\mu}_t, \hat{\beta}_t) : \operatorname{rdepth}\left((\hat{\mu}_t, \hat{\beta}_t), \mathbf{y}_t\right)\right\},$$

where $\operatorname{rdepth}((\hat{\mu}_t, \hat{\beta}_t), \mathbf{y}_t)$ is the *regression depth* of a fit $(\hat{\mu}_t, \hat{\beta}_t)$ to a sample $\mathbf{y}_t$:

$$\operatorname{rdepth}\big((\hat{\mu}_t, \hat{\beta}_t), \mathbf{y}_t\big) = \min_{i \in I}\big\{\min\big\{L^+(i) + R^-(i), R^+(i) + L^-(i)\big\}\big\},$$

$$L^+(i) = \#\{j \in I_{\text{left}} : \mathfrak{r}_{t+j} \geq 0\},$$

$$R^-(i) = \#\{j \in I_{\text{right}} : \mathfrak{r}_{t+j} < 0\}$$

with $I_{\text{left}} = \{-w, \ldots, i\}$ and $I_{\text{right}} = \{i+1, \ldots, w\}$ for the delayed approach and $I_{\text{left}} = \{-n+1, \ldots, -n+i\}$ and $I_{\text{right}} = \{-n+i+1, \ldots, 0\}$ for the online approach. $L^-(i)$ and $R^+(i)$ are defined in the same way. In case of more than one fit with maximal regression depth, the deepest regression estimate is the average of these fits. The DR is computationally less expensive than LMS and *Least Trimmed Squares* (LTS, Rousseeuw 1984) regression, but it has a fsbp of only $1/3$ in case of equally spaced design points (Gather et al. 2006). For more details regarding depth statistics, see the contribution by Mosler, Chap. 2.

The *Repeated Median* (RM, Siegel 1982) regression estimator $(\hat{\mu}_t^{\text{RM}}, \hat{\beta}_t^{\text{RM}})$ is given by

$$\hat{\beta}_t^{\text{RM}} = \operatorname*{med}_{i \in I}\left\{\operatorname*{med}_{j \in I \setminus \{i\}} \frac{y_{t+i} - y_{t+j}}{i - j}\right\},$$

$$\hat{\mu}_t^{\text{RM}} = \operatorname*{med}_{i \in I}\big\{y_{t+i} - i\hat{\beta}_t^{\text{RM}}\big\}.$$

The RM is highly robust with a fsbp of $\lfloor n/2 \rfloor / n$. A fast update-algorithm allows calculation of the RM in linear time (Bernholt and Fried 2003).

Davies et al. (2004) examine delayed L_1, RM and LMS regression with respect to their suitability for real time signal extraction. They compare the robustness, the computing time, the efficiency for Gaussian noise and the tracking of level shifts and trend changes, and find the RM to deliver the best overall performance. It offers higher robustness than the L_1 and yields smoother signal extractions and faster computation than the LMS. However, the LMS traces level shifts and trend changes better than the RM.

Gather et al. (2006) examine DR, RM, LMS, and LTS regression for online signal extraction. Again, the RM shows the best overall performance w.r.t. robustness, computing time and efficiency given various data situations.

The RM is obviously a good candidate for real time signal extraction. It is Lipschitz-continuous in case of equidistant design points, which means that small changes in the data do not lead to large changes of the signal estimate. Furthermore, the RM slope estimator is unbiased and Fisher-consistent for a fixed design and symmetrically distributed errors (Siegel 1982). In addition, the RM is regression and scale equivariant. For finite Gaussian samples, the estimator $\hat{\mu}_t^{\text{RM}}$ shows a quite large efficiency of more than 60 % compared to the least squares estimator (Gather et al. 2006), and its asymptotic efficiency is 63.7 % (Hössjer et al. 1995).

There are several extensions and modifications of the RM, mainly for improving its efficiency and its ability to trace level shifts. These are, amongst others, the

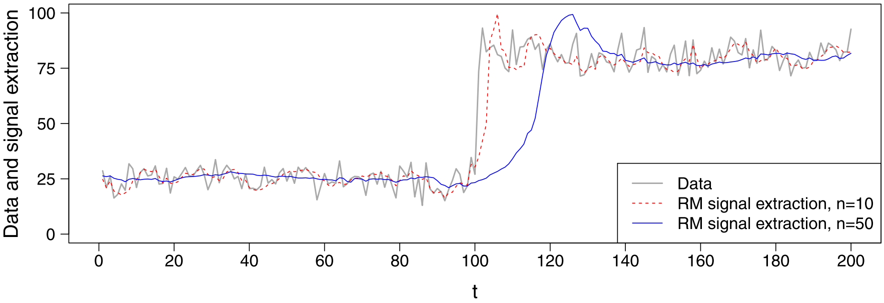


Fig. 12.2 Bias-variance-dilemma for the choice of the window width: a large value of n leads to smooth signal extractions, and a small n to signal extractions which are close to the data

Predictive Repeated Median Hybrid filter (PRMH, Fried et al. 2006), *Weighted Repeated Median* regression (WRM, Fried et al. 2007) and *Trimmed Repeated Median* (TRM) regression (Bernholt et al. 2006). Furthermore, Fried (2004) proposes an extension of the RM which includes rules for outlier detection and shift preservation. Like for all localized signal extraction procedures, the RM is affected by the well-known bias-variance-dilemma for the choice of the window width n. As long as the data stream shows a stable trend, a large n should be used to obtain smooth signal extractions. However, when a sudden change occurs in the data (e.g. a level shift), the window width n should be chosen small to ensure that the signal estimate traces the change with high accuracy. Figure 12.2 illustrates this dilemma for the choice of the window width n. Some of the aforementioned RM-based signal filters improve the tracking of level shifts, but all use a fixed window width specified by the analyst. This motivates the data-adaptive choice of the window width, meaning that n should be adjusted to the current data situation at each time point t.

12.4 RM-Based Filters with Data-Adaptive Width Selection

Gather and Fried (2004) propose a window width adaption approach for the delayed RM filter. Their idea is adopted by Schettlinger et al. (2010) who develop an RM-based online filter with data-adaptive width selection, the *adaptive online RM* (aoRM) filter.

12.4.1 The aoRM

The aoRM selects the window width according to the current data situation at each point in time. The window width at time t is denoted by n_t in the following. The aoRM uses a goodness-of-fit test to decide whether the window width should be adapted. In short, the aoRM algorithm works as follows: Given a time point t and a window width n_t, an RM regression is performed in the time window $\{t - n_t + 1, \ldots, n_t\}$. Then the aoRM uses a test to decide whether the RM fit is adequate or not. If the RM fit is not assessed to be adequate, the window width n_t is decreased.

The test applied by the aoRM uses the fact that an RM regression results in an equal number of positive and negative residuals. The basic idea of the test is that an RM fit is adequate only if the balance of the residual signs is given for any subset of design points. Given an RM fit in a time window $\{t - n_t + 1, \ldots, t\}$, the aoRM tests the null hypothesis that the median of the distribution of the $m \leq \lfloor n_t/2 \rfloor$ rightmost RM residual signs is zero, against the alternative that the median is not zero. Let $\{\mathfrak{r}_{t-n+i}\}_{i \in \{1,\ldots,n\}}$ denote the residuals of the RM fit in a time window of width n (for simplicity, we omit the time index t here). The aoRM test statistic at time t is then the absolute sum of the signs of the $m \leq \lfloor n/2 \rfloor$ rightmost RM residuals:

$$T_t^{\text{aoRM}} = \left| \sum_{i \in J} \text{sign}(\mathfrak{r}_{t-n+i}) \right|, \quad J = \{m+1, \ldots, n\}, \tag{12.3}$$

where sign$(\cdot)$ denotes the sign function:

$$\text{sign}(\omega) = \begin{cases} -1, & \text{if } \omega < 0, \\ 0, & \text{if } \omega = 0, \\ 1, & \text{if } \omega > 0. \end{cases}$$

The null hypothesis is rejected if the test statistic exceeds a critical value which depends on n, m and the significance level α. The critical values are obtained by Monte Carlo simulations for small n and m; for large values of n and m, Schettlinger et al. (2010) use quantiles of a hypergeometric distribution.

The aoRM test is applied at each time t to find an adequate window width n_t. Before we explain the aoRM algorithm we briefly discuss the input arguments that must be chosen by the user (for more details, see Schettlinger et al. 2010):

$n_{\min}$	minimum window width,
$n_{\max}$	maximum window width,
m	size of the sub-window $\{t - m + 1, \ldots, t\}$ used for testing,
α	significance level of the test.

The minimum and maximum window width ensure that $n_t \in \{n_{\min}, \ldots, n_{\max}\} \subset \mathbb{N}$. The minimum width guarantees robustness against a certain number of outliers. For example, if it is justifiable to assume that up to 10 outliers occur in a patch, $n_{\min}$ should be greater than 21, due to the fsbp of the RM. The maximum window width limits the computing time. On basis of a simulation study, Schettlinger et al. (2010) suggest to choose the size of the sub-window m such that $m \leq n_{\min}/2$. Furthermore, they find the aoRM to show good performance using the significance level $\alpha = 0.1$.

The following algorithm starts as soon as $n_{\min}$ observations are given, i.e. at time $t = n_t = n_{\min}$:

1. Perform an RM regression in the time window $\{t - n_t + 1, \ldots, t\}$ and obtain $\hat{\mu}_t^{\text{RM}}$.
2. If $n_t = n_{\min}$, go to step 4.

3. Perform test; if H_0 is rejected, set n_t to $n_t - 1$ and go back to step 1.
4. Store $\hat{\mu}_t^{\mathrm{RM}}$ as aoRM signal estimate for time t.
5. Update window: set n_{t+1} to $\min\{n_t + 1, n_{\max}\}$;
   Update index: set $t + 1$ to t.

At step 3 the aoRM tests the adequacy of the RM fit. If the null hypothesis is rejected, i.e., if the fit is assessed to be inadequate, the algorithm sets the window width n_t to $n_t - 1$, i.e., the oldest/leftmost observation is excluded from the sample. A new RM regression is then performed on the decreased sample (step 1), and it is tested again until either the test cannot reject the null hypothesis or n_t equals $n_{\min}$ (step 2). In either case, $\hat{\mu}_t^{RM}$ is stored as aoRM signal estimate for time t (step 4). At step 5, the aoRM updates the time window for the next time point $t + 1$. This is done by including the incoming new observation y_{t+1} into the window sample, so that it grows up to size $n_t + 1$. If $n_t + 1 > n_{\max}$, the oldest/leftmost window observation is excluded from the window, so that $n_{t+1} = n_{\max}$. The update step is finished by setting the index $t + 1$ to t.

The aoRM decreases the window width by one when the data in the time window cannot be assumed to have a linear structure. In other words, the aoRM reacts to changes of the signal. Borowski and Fried (2011) propose an alternative RM-based filter with data-adaptive width selection. Their *Slope Comparing Adaptive Repeated Median* (SCARM) uses a different test procedure and a different rule for adapting the window width.

12.4.2 The SCARM

At each time t, the SCARM tests the null hypothesis of a linear underlying signal in the time window of size n_t, against the alternative that a level shift or a trend change occurs. To test this null hypothesis, the whole window $\{t - n_t + 1, \dots, t\}$ is divided into two separate windows, a left-hand window $\{t - n_t + 1, \dots, t - n_t + \ell_t\}$ and a right-hand window $\{t - r + 1, \dots, t\}$. The right-hand width r is fixed whereas the left-hand width ℓ_t changes as n_t changes, where $\ell_t = n_t - r$. The SCARM test statistic is then

$$T_t^{\mathrm{SCARM}} := \frac{D_t}{\sqrt{\mathrm{Var}(D_t)}}, \qquad D_t := \hat{\beta}_t^{\mathrm{left}} - \hat{\beta}_t^{\mathrm{right}},$$

where $\hat{\beta}_t^{\mathrm{left}}$ and $\hat{\beta}_t^{\mathrm{right}}$ denote the RM slopes estimated from the left-hand and right-hand window, respectively. Borowski and Fried (2011) propose an approach to estimate $\mathrm{Var}(D_t)$ which prevents *masking effects* induced by level shifts and improves the power of the test. (Here masking means that the estimate of $\mathrm{Var}(D_t)$ is too large for the test to detect an existing signal change.) The null hypothesis is rejected if $|T_t^{\mathrm{SCARM}}|$ exceeds the $1 - (\alpha/2)$-quantile of a t-distribution with f degrees of freedom, where f depends on ℓ_t and r. A rejection of H_0 means that a level shift and/or trend change is assumed, so that n_t is not adequate but must be decreased.

The SCARM requires the choice of the following input arguments:

r	fixed width of the right-hand window r,
$\ell_{\min}$	minimum left-hand width,
$n_{\min}$	minimum window width,
$n_{\max}$	maximum window width,
α	significance level of the test.

Like for the aoRM, $n_{\min}$ and $n_{\max}$ ensure that $n_t \in \{n_{\min}, \dots, n_{\max}\} \subset \mathbb{N}$. The minimum left-hand width $\ell_{\min}$ ensures that the test is only performed if $n_t = \ell_t + r \geq \ell_{\min} + r$, i.e., if the window contains enough observations for a meaningful testing. Borowski and Fried (2011) suggest to choose identical values for r and $\ell_{\min}$, that are greater than three times the length of the largest outlier patch expected. The reason is that the RM is considerably biased if the proportion of aberrant observations is around $1/3$. This can be seen in Fig. 12.2, where the RM signal extractions react to the level shift with a delay of approximately $n/3$ time points. Therefore, if the SCARM test detects a signal change at time t, it can be assumed that the change happened around time $t - r/3$. Hence, it is recommended to choose $n_{\min} = r/3$.

The SCARM starts the signal extraction as soon as $n_{\min}$ observations are given, i.e. at time $t = n_t = n_{\min}$. The algorithm is then as follows:

1. If $n_t < \ell_{\min} + r$, go to step 3.
2. Perform SCARM test; if H_0 is rejected, set n_t to $n_{\min}$.
3. Perform RM regression in time window $\{t - n_t + 1, \dots, t\}$ and store $\hat{\mu}_t^{\mathrm{RM}}$ as SCARM signal estimate for time t.
4. Update window: set n_{t+1} to $\min\{n_t + 1, n_{\max}\}$;
   Update index: set $t + 1$ to t.

The SCARM and aoRM use different principles to decrease the window width. The aoRM aims at the largest window size for which H_0 cannot be rejected. In contrast, the SCARM searches for changes of the signal and sets n_t to the minimal value $n_{\min}$ whenever a change is detected to achieve that the RM depicts the change with high accuracy.

The aoRM and SCARM finally estimate the signal by means of online RM regression, meaning that their performance is strongly related to the performance of the test procedure. Hence, Borowski and Fried (2011) compare the two filters by focusing on the properties of the specific tests, and find that the test of the SCARM outperforms the aoRM test. The tests show comparable robustness properties, but the SCARM test offers larger power in particular for small window samples and also detects changes much faster than the aoRM in case of Gaussian noise. Thus, the SCARM traces level shifts and trend changes more accurately than the aoRM.

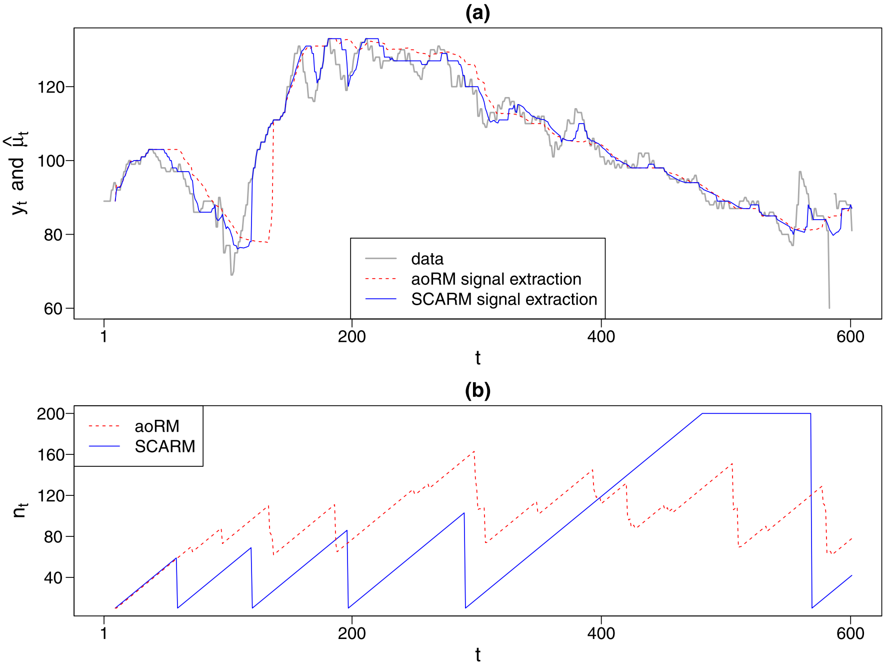


Fig. 12.3 (**a**) Time series and signal extraction by aoRM and SCARM; (**b**) adapted window widths

12.4.3 Application

We apply the aoRM and SCARM retrospectively to a time series (y_t), $t = 1, \ldots, 600$, of blood pressure measurements using the R functions *adore.filter* (*ad*aptive *o*nline *re*peated median filter) and *scarm.filter* from the R package *robfilter* (Fried et al. 2012). The following input arguments are used: $n_{\min} = 10$, $n_{\max} = 200$, $m = 30 = r = \ell_{\min}$, and $\alpha = 0.01$. The time series (y_t) and the signal extraction of the aoRM and SCARM are shown in Fig. 12.3(a); the lower part (b) of Fig. 12.3 shows the adapted window widths. Apparently, it is difficult if not impossible to reproduce the course of the data time series (y_t) by the time series (n_t) of the widths adapted by the aoRM (red, dashed). This fact has already been remarked by Schettlinger (2009, Chap. 3.3.3). In contrast, the window widths of the SCARM (blue, solid) do allow a conclusion about the course of the data time series (y_t). When the time series (y_t) show a trend change or level shift, the SCARM sets the window width n_t down to its minimum value $n_{\min}$. Afterwards, the SCARM increases n_t successively until the next signal change is detected. In our view, the SCARM detects all relevant signal changes here, and no type I error happens. As a consequence, the SCARM signal estimates (blue, solid) are close to the data when there are signal changes and is smooth when the course of the time series (y_t) is stable. In contrast, the aoRM does not fulfill both aims simultaneously: The aoRM signal extraction (red, dashed) is smooth, but signal changes are not traced exactly, cf. the distinct signal change around time $t = 100$ for instance.

12.5 RM-Based Filters for Multivariate Time Series

Signal extraction from a multivariate data stream could be achieved by applying a univariate filter to each component of the multivariate stream. However, this approach does not consider the dependencies between the measured variables. We therefore present two recently developed multivariate RM-based filters, which include the information given by the cross-dependence structure.

The general assumption (12.1) of an underlying signal which is interfered by noise and outliers can be extended to multivariate time series (y_t) with $y_t = [y_t(1), \dots, y_t(K)]'$:

$$Y_t = \mu_t + \varepsilon_t + \eta_t, \quad t \in \mathbb{Z}, \tag{12.4}$$

where $\mu_t = [\mu_t(1), \dots, \mu_t(K)]'$ is the K-variate signal vector at time t and $\varepsilon_t = [\varepsilon_t(1), \dots, \varepsilon_t(K)]'$ the multivariate noise generating process with $E[\varepsilon_t(k)] = 0$ for all $k = 1, \dots, K$ and smoothly varying covariance matrix $\mathrm{Cov}[\varepsilon_t] = \mathbf{\Sigma}_t$. The covariance matrix can be non-diagonal, i.e., possibly $\mathrm{Cov}[\varepsilon_t(i), \varepsilon_t(j)] = \sigma_t(i, j) \neq 0$ for some $i \neq j$, to allow for correlations between error components. Like for the univariate case, the outlier generating mechanism $\eta_t \in \mathbb{R}^K$ produces impulsive, spiky noise but is zero most of the time.

Referring to the working model (12.4), Lanius and Gather (2010) transfer the local linearity assumption of the delayed signal extraction approach (12.2) to the multivariate case: They assume that the K-variate signal can be approximated locally by K regression lines in a time window $\{t - w, \dots, t, \dots, t + w\}$ of width $n = 2w + 1$. The central level of the k-th regression line is then the k-th component of the signal estimation vector $\hat{\mu}_t$. Lanius and Gather propose the multivariate *Trimmed Repeated Median-Least Squares* (TRM-LS) regression which is an enhancement of the univariate TRM-LS by Bernholt et al. (2006), mentioned in Sect. 12.3.

12.5.1 *The TRM-LS*

Given a window sample $(y_{t-w}, \dots, y_t, \dots, y_{t+w}) \in \mathbb{R}^{K \times n}$, the algorithm of the multivariate TRM-LS regression is as follows:

1. Perform an RM regression for each of the K univariate window samples $(y_{t-w}(k), \dots, y_t(k), \dots, y_{t+w}(k))$, $k = 1, \dots, K$.
2. Regard the resulting RM residuals as a $(K \times n)$-sample $(\mathfrak{r}_{t-w}, \dots, \mathfrak{r}_t, \dots, \mathfrak{r}_{t+w})$, where $\mathfrak{r}_{t+i} = [\mathfrak{r}_{t+i}(1), \dots, \mathfrak{r}_{t+i}(K)]'$, $i = -w, \dots, w$, and estimate the local error covariance matrix Σ_t on this residual sample using a robust covariance estimator.

3. Use the estimate $\hat{\Sigma}_t$ to detect residual vectors which are outliers w.r.t. the local covariance structure, i.e., residual vectors $\mathfrak{r}_{t+i}$ with
$$\mathfrak{r}'_{t+i}\hat{\Sigma}_t^{-1}\mathfrak{r}_{t+i} > d,$$
where $d > 0$ is a suitable upper bound. Remove those observation vectors y_{t+i} from the sample, which belong to an outlying residual vector.
4. Perform a multivariate Least Squares regression on the trimmed window sample.

Lanius and Gather (2010) use the robust and fast computable *orthogonalized Gnanadesikan-Kettenring estimator* (Gnanadesikan and Kettenring 1972; Maronna and Zamar 2002) to estimate the covariance matrix Σ_t. A typical choice for the upper trimming bound d in step 4 is $d = \chi^2_K(0.95)$, the 0.95-quantile of a χ^2-distribution with K degrees of freedom. For more details regarding the estimation of Σ_t and the choice of d, see Lanius and Gather (2010).

The TRM-LS filter offers robustness against outliers in one or more components as well as against outliers w.r.t. the local dependence structure. Such outliers would possibly not be detected by univariate methods. Furthermore, the TRM-LS offers high efficiency at Gaussian samples since LS regression is used for the final signal estimate. However, the TRM-LS is a delayed signal filter and uses a fixed window width. Hence, Borowski et al. (2009) transfer the delayed TRM-LS to the online case and combine it with the univariate aoRM. Their *adaptive online TRM-LS* (aoTRM-LS) is a multivariate online filter with adaptive width selection.

12.5.2 The aoTRM-LS

Like the TRM-LS, the aoTRM-LS is based on the general assumption (12.4). However, like the aoRM, the aoTRM-LS uses the time window $\{t - n_t + 1, \ldots, t\}$, choosing the width n_t at each time point with regard to the current data situation and delivering estimates of the signal vector at the most recent time point t. The aoTRM-LS requires the prior specification of the same input arguments as the aoRM, namely $n_{\min}$, $n_{\max}$, m and α. Signal extraction is started as soon as $n_{\min}$ observations are present, i.e. at time $t = n_t = n_{\min}$. The algorithm of the aoTRM-LS is then as follows:

1. Apply the aoRM window width adaption procedure separately to each of the K univariate window samples $(y_{t-n_t+1}(k), \ldots, y_t(k))$ to obtain K individual adapted widths $n_t(k)$, $k = 1, \ldots, K$.
2. Set the overall window width $n_t := \min_k\{n_t(k)\}$.
3. Estimate the signal vector μ_t from the multivariate sample $(y_{t-n_t+1}, \ldots, y_t) \in \mathbb{R}^{K\times n_t}$ by the online TRM-LS.
4. Update window: set n_{t+1} to $\min\{n_t + 1, n_{\max}\}$;
   Update index: set $t + 1$ to t.

The overall window width n_t is chosen as the minimum of the individual window widths $n_t(k)$ in order not to violate the assumption of an underlying linear signal (step 2). This necessity implies that n_t is chosen as $n_{\min}$ most of the time if K gets large. To overcome this problem, Borowski et al. (2009) suggest to apply the aoTRM-LS *block-wise*. That is, the user divides the univariate components of the multivariate time series into disjoint sets (the blocks), and the filter is then applied separately to each block. Since the blocks are disjoint, a small individual window width $n_t(k)$ in one block does not affect the width in another block.

12.6 Conclusions

Real time signal extraction from noisy and outlier-contaminated data stream time series can be achieved by applying robust location or regression estimators to moving time windows containing the n most recent observations. The (double window) modified trimmed mean combines the robustness of the running median with the efficiency of the running mean. Linear median hybrid filters take the median of linear subfilter outcomes. Since the subfilters can be chosen w.r.t. the assumed local polynomial degree of the underlying signal, this class of filters is very flexible. These location-based filters are generally designed for locally constant underlying signals. For time series that exhibit trends, signal filters based on local fitting of regression lines are generally the better choice.

Regression filters can be applied in a delayed or in an online fashion. Since the online approach estimates the signal without any time delay (except computing time), it is recommended for real time applications. The simple robust regression estimators presented here are the L_1, Least Median of Squares (LMS) regression, the Deepest Regression (DR), and the Repeated Median (RM) regression. The RM has turned out to deliver the best overall performance w.r.t. robustness, computing time, and efficiency, which is why there exist several extensions and modifications of the RM.

The RM is a good candidate for real time signal extraction, but as any other localized filtering procedure it cannot escape the bias-variance-dilemma for the choice of the window width. The adaptive online RM (aoRM) and the Slope Comparing Adaptive RM (SCARM) tackle this problem by choosing the window width at each time t with respect to the current data situation. Both filters use a test to decide whether the window width should be adapted. The test of the aoRM examines the balance of the residual signs, whereas the SCARM test compares the RM slopes within separate time windows. The SCARM is recommended in case of Gaussian noise with outliers.

There exist also RM-based filters for multivariate time series which make use of the local cross-dependencies. The Trimmed Repeated Median-Least Squares (TRM-LS) procedure first fits K RM regression lines, one for each component of the K-variate window sample. The local error covariance matrix is then estimated on the RM residuals and used to detect outlying observation vectors. The signal is finally estimated on the trimmed window sample by means of Least Squares regression. The *adaptive online TRM-LS* (aoTRM-LS) combines the advantages of the adaptive univariate aoRM and the non-adaptive multivariate TRM-LS.

The presented robust regression filters, including the aoRM, SCARM and aoTRM-LS, are provided in the R package *robfilter* (Fried et al. 2012). The R-functions of the aoRM and SCARM are denoted as *adore.filter* (*ad*aptive *o*nline *re*peated median filter) and *scarm.filter*. The R-function of the multivariate aoTRM-LS is denoted as *madore.filter* (*m*ultivariate *ad*aptive *o*nline *re*peated median filter).

Acknowledgements We gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (SFB475 "Reduction of Complexity in Multivariate Data Structures", 1997–2009, and SFB823 "Statistical modelling of nonlinear dynamic processes").

References

Astola, J., Heinonen, P., & Neuvo, Y. (1989). Linear median hybrid filters. *IEEE Transactions on Circuits and Systems*, *36*, 1430–1438.

Bernholt, T., & Fried, R. (2003). Computing the update of the repeated median regression line in linear time. *Information Processing Letters*, *88*(3), 111–117.

Bernholt, T., Fried, R., Gather, U., & Wegener, I. (2006). Modified repeated median filters. *Statistics and Computing*, *16*, 177–192.

Borowski, M., & Fried, R. (2011). *Robust repeated median regression in moving windows with data-adaptive width selection*. Discussion paper 28/2011, SFB 823, Technische Universität Dortmund.

Borowski, M., Schettlinger, K., & Gather, U. (2009). Multivariate real time signal processing by a robust adaptive regression filter. *Communications in Statistics. Simulation and Computation*, *38*(2), 426–440.

Davies, P. L., Fried, R., & Gather, U. (2004). Robust signal extraction for on-line monitoring data. *Journal of Statistical Planning and Inference*, *122*, 65–78.

Davies, P. L., & Gather, U. (2005). Breakdown and groups. *The Annals of Statistics*, *33*, 977–1035.

Donoho, D. L., & Huber, P. J. (1983). The notion of breakdown point. In P. J. Bickel, K. A. Doksum, & J. L. Hodges (Eds.), *Festschrift for Erich Lehmann* (pp. 157–184). Belmont: Wadsworth.

Fried, R. (2004). Robust filtering of time series with trends. *Journal of Nonparametric Statistics*, *16*, 313–328.

Fried, R., Bernholt, T., & Gather, U. (2006). Repeated median and hybrid filters. *Computational Statistics & Data Analysis*, *50*, 2313–2338.

Fried, R., Einbeck, J., & Gather, U. (2007). Weighted repeated median smoothing and filtering. *Journal of the American Statistical Association*, *102*, 1300–1308.

Fried, R., Schettlinger, K., & Borowski, M. (2012). *robfilter: robust time series filters*. R package version 4.0. http://CRAN.R-project.org/package=robfilter. Cited in October 4, 2012.

Gather, U., & Fried, R. (2004). Methods and algorithms for robust filtering. In J. Antoch (Ed.), *Proceedings in computational statistics COMPSTAT 2004* (pp. 159–170). Heidelberg: Physika-Verlag.

Gather, U., Schettlinger, K., & Fried, R. (2006). Online signal extraction by robust linear regression. *Computational Statistics*, *21*, 33–51.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, *28*(1), 81–124.

Hampel, F. R. (1975). Beyond location parameters: robust concepts and methods. *Bulletin de L'Institut International de Statistique*, *46*, 375–382.

Heinonen, P., & Neuvo, Y. (1987). FIR-median hybrid filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *35*, 832–838.

Heinonen, P., & Neuvo, Y. (1988). FIR-median hybrid filters with predictive FIR substructures. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *36*, 892–899.

Hössjer, O., Rousseeuw, P. J., & Ruts, I. (1995). The repeated median intercept estimator: influence function and asymptotic normality. *Journal of Multivariate Analysis*, *52*, 45–72.

Lanius, V., & Gather, U. (2010). Robust online signal extraction from multivariate time series. *Computational Statistics & Data Analysis*, *54*, 966–975.

Lee, Y., & Kassam, S. (1985). Generalized median filtering and related nonlinear filtering techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *33*, 672–683.

Maronna, R. A., & Zamar, R. H. (2002). Robust estimates of location and dispersion for high dimensional datasets. *Technometrics*, *44*, 307–317.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*, 871–880.

Rousseeuw, P. J., & Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association*, *99*, 388–402.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Schettlinger, K. (2009). *Signal and variability extraction for online monitoring in intensive care*. Ph.D. Thesis, Fakultät Statistik, Technische Universiät Dortmund.

Schettlinger, K., Fried, R., & Gather, U. (2010). Real time signal processing by adaptive repeated median filters. *International Journal of Adaptive Control and Signal Processing*, *24*, 346–362.

Siegel, A. F. (1982). Robust regression using repeated medians. *Biometrika*, *69*(1), 242–244.

Sposito, V. A. (1990). Some properties of L$_p$ estimators. In K. D. Lawrence & J. L. Arthur (Eds.), *Robust regression—analysis and applications* (pp. 23–58). New York: Dekker.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.

Wichman, R., Astola, J. T., Heinonen, P. J., & Neuvo, Y. A. (1990). FIR-Median hybrid filters with excellent transient response in noisy conditions. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *38*, 2108–2117.

Chapter 13
Robustness in Time Series: Robust Frequency Domain Analysis

Bernhard Spangl and Rudolf Dutter

13.1 Introduction

The spectral density function is commonly used to analyze time series. Areas of application are signal processing (cf. Thomson 1994), geophysics (cf. Chave et al. 1987; Jones and Hollinger 1997) and medicine.

This chapter is motivated by the frequency-domain analysis of short-term heart rate variability (HRV) recordings. This is a non-invasive method which has been increasingly used in medicine (cf. Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology 1996; Howorka et al. 1997, 1998; Hartikainen et al. 1998; Pumprla et al. 2002). To access the variability of heart rate in the frequency domain the spectral density function of the tachogram is estimated. The tachogram is the series of time intervals between consecutive heart beats. These time intervals are also called *R*–*R*-intervals, i.e., the periods between an *R*-peak and the next *R*-peak in an electrocardiogram. The intervals normally have a duration of about 750 ms corresponding to a heart rate of 80 beats per minute. In the tachogram (an example is displayed in Fig. 13.1), outlying observations can be caused by ventricular ectopic beats and other artifacts (cf. Hartikainen et al. 1998). Ectopic beats are usually premature and produce a very short *R*–*R*-interval followed by a compensatory delay and therefore a prolonged *R*–*R*-interval. Typical tachogram patterns caused by ectopic beats can be seen in Fig. 13.1 around heart beat number 90 and 1090. Correspondingly, missed beats result in erroneously prolonged *R*–*R*-intervals (sum of two consecutive *R*–*R*-intervals). Typical patterns caused by missed beats are visible in Fig. 13.1 around beat number 730.

B. Spangl (✉)
University of Natural Resources and Life Sciences, Gregor-Mendel-Str. 33, 1180 Vienna, Austria
e-mail: bernhard.spangl@boku.ac.at

R. Dutter
Vienna University of Technology, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria
e-mail: R.Dutter@tuwien.ac.at

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_13, © Springer-Verlag Berlin Heidelberg 2013

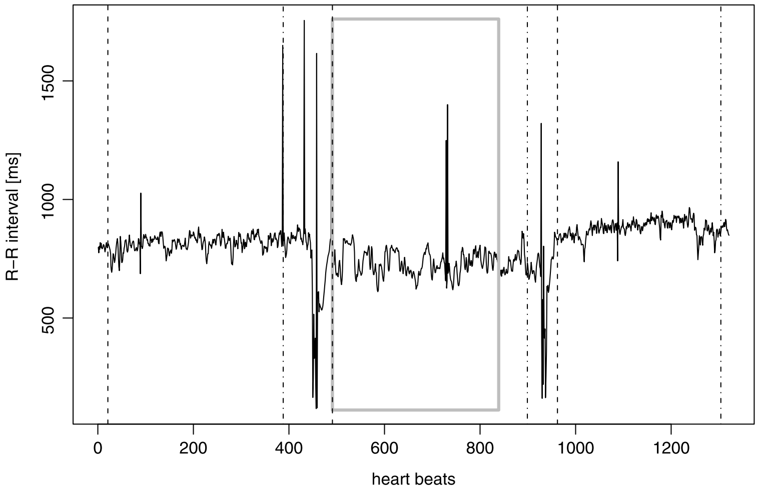


Fig. 13.1 Tachogram of 1321 consecutive heart beats

These outlying tachogram measurements affect the spectral analysis of heart rate variability if classical spectral density estimators, which are sensitive to outliers, are used. For details see Kleiner et al. (1979) or Martin and Thomson (1982). Therefore, we aim to access the heart rate variability by estimating the spectral density function of the tachogram series using robust methods that are insensitive to outlying tachogram values caused by ectopic beats or other artifacts. Furthermore, as ectopic or missing beats do not affect successive heart beats, the additive outlier (AO) model (cf. Denby and Martin 1979; Fox 1972) seems to be an appropriate model when analyzing heart rate variability data. The AO model consists of a stationary core process, x_t, to which occasional outliers have been added. The observed process y_t is said to have additive outliers if it is defined by

$$y_t = x_t + v_t, \tag{13.1}$$

where the contaminations v_t are independently and identically distributed according to a contaminated normal distribution with a degenerate central component, i.e.,

$$\mathcal{CN}\big(\gamma, 0, \sigma^2\big) = (1-\gamma)\mathcal{N}(0,0) + \gamma\mathcal{N}\big(0, \sigma^2\big), \tag{13.2}$$

where γ is small. Let $\mathcal{N}(\mu, \sigma^2)$ denote the normal distribution with mean μ and variance σ^2. Hence, the core process x_t is observed with probability $1-\gamma$ whereas the core process plus a disturbance v_t is observed with probability γ. x_t and v_t are also assumed to be independent. Here, especially if caused by ectopic beats, additive outliers can be positive as well as negative. Other types of outliers are described in detail in the contribution by Galeano and Peña, Chap. 15.

We do not compute the spectral density function of the entire tachogram, but calculate several estimates within overlapping windows (cf. Pumprla et al. 2002). This is to ensure stationarity in each window and to deal with signals whose frequency

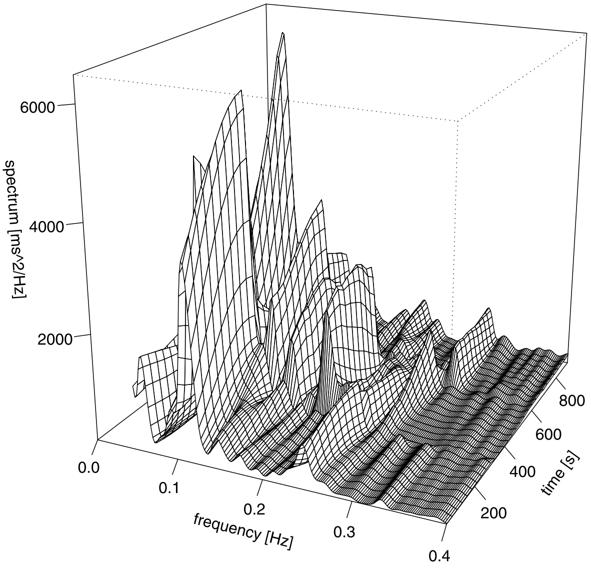


Fig. 13.2 Robust dynamic Fourier analysis of the original short-term HRV data displayed in Fig. 13.1

content changes over time. The result of the so called dynamic Fourier analysis applied to the tachogram series plotted in Fig. 13.1 is displayed in Fig. 13.2. Each slice parallel to the frequency-spectrum plane in Fig. 13.2 represents the spectral density estimate of the corresponding time window.

A high variability in heart rate indicates good adaptability, implying a healthy person with well functioning autonomic control mechanisms. Conversely, lower variability is often an indicator of abnormal and insufficient adaptability of the autonomic nervous system.

In the following, the problem of estimating the spectral density function robustly is considered. In Sect. 13.2, several methods of classical and robust spectral density estimation are described. At the end of Sect. 13.2, an outline of our small simulation study is given and the results are presented. An application of robust spectral density estimation to the analysis of heart rate variability is given in Sect. 13.3 and some general remarks follow in Sect. 13.4.

13.2 Methods

13.2.1 Classical Spectral Density Estimation

13.2.1.1 The Spectral Representation Theorem

A stochastic process $\{z(f) : f \in [-1/2, 1/2]\}$ with random variables $z(f) : \Omega \to \mathbb{C}^p$ is called a *process of orthogonal increments* if the following conditions are satisfied:

1. $z(-1/2) = 0$ a.e. and $z(1/2) = x_0$ a.e.,
2. $\lim_{\varepsilon\downarrow 0} z(f + \varepsilon) = z(f)$ for $f \in [-1/2, 1/2)$ (right continuity)
3. $E[z(f)^* z(f)] < \infty$ for all $f \in [-1/2, 1/2]$
4. $E[(z(f_4) - z(f_3))(z(f_2) - z(f_1))^*] = 0$ for all $f_1 < f_2 \leq f_3 < f_4$,

where the notation $\text{l.i.m}_{k\to\infty} x_k = x_0$ means that the limit is understood in the mean squares sense.

More precisely, a sequence of random variables $\{x_k\}_{k\in\mathbb{N}}$ is said to converge to x_0 in *mean squares sense* if

$$E\big[x_0^* x_0\big] < \infty$$

and

$$\lim_{k\to\infty} E\big[(x_k - x_0)^*(x_k - x_0)\big] = 0$$

holds.

Moreover, we note that $\{z(f) : f \in [-1/2, 1/2]\}$ is a stochastic process with a continuous index set $[-1/2, 1/2]$ and we will interpret these indices f not as time points but as frequencies.

For every second-order stationary process $\{x_t\}$ there exists a process $\{z(f) : f \in [-1/2, 1/2]\}$ with orthogonal increments such that

$$x_t = \int_{-1/2}^{1/2} e^{i2\pi f t}\, dz(f)$$

holds. The process $\{z(f)\}$ is a.e. uniquely determined by $\{x_t\}$.

The process $\{z(f)\}$ defines a function $F : [-1/2, 1/2] \to \mathbb{C}^{n\times n}$ by $F(f) = E[z(f)z(f)^*]$ where the following relations hold:

$$\begin{aligned} F(-1/2) &= 0 \\ F(1/2) &\geq 0 \\ F(f_2) - F(f_1) &= E\big[\big(z(f_2) - z(f_1)\big)\big(z(f_2) - z(f_1)\big)^*\big] \quad \text{for } f_1 \leq f_2. \end{aligned}$$

Thus $F(\cdot)$ is a non-decreasing right continuous function, where non-decreasing means that the difference $F(f_2) - F(f_1)$ is a non-negative definite matrix for all $f_1 \leq f_2$.

If $\{z(f)\}$ is the orthogonal increment process corresponding to $\{x_t\}$ the function $F(\cdot)$ is called the *spectral distribution function* of $\{x_t\}$. If there exists a function $S : [-1/2, 1/2] \to \mathbb{C}^{n\times n}$ such that

$$F(f) = \int_{-1/2}^{f} S(\nu)\, d\nu,$$

where ν denotes the Lebesgue measure, then $S(\cdot)$ is called the *spectral density function* of $\{x_t\}$. Other commonly used terms for $S(\cdot)$ are *spectral density*, *spectrum* or *power spectrum*.

One condition to ensure the existence of the spectral density function is if the autocovariance function $\{\gamma(h)\}$ is absolutely summable, i.e.,

$$\sum_{h=-\infty}^{\infty} \left|\gamma(h)\right| < \infty.$$

Then the spectral distribution function is absolutely continuous with $dF(f) = S(f)\,df$. Under this condition, the autocovariance function at lags h, $\gamma(h)$, $h \in \mathbb{Z}$, are the Fourier coefficients of $S(\cdot)$ and thus we can represent $S(f)$ as

$$S(f) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-i2\pi f h}.$$

13.2.1.2 Nonparametric Estimation

The nonparametric estimation of the spectral density function is based on smoothing the periodogram.

Let $\{x_t, t = 1, \ldots, n\}$ denote the observed process which is assumed to be second-order stationary and to have zero mean. Further, suppose that the time intervals between two consecutive observations are equally spaced with duration Δt. Then the *periodogram* is defined as follows:

$$\hat{S}^{(p)}(f) = \frac{\Delta t}{n} \left| \sum_{t=1}^{n} x_t e^{-i2\pi f t \Delta t} \right|^2 = \Delta t \sum_{h=-(n-1)}^{(n-1)} \hat{\gamma}_x(h) e^{-i2\pi f h \Delta t},$$

where $\{\hat{\gamma}_x(h)\}$ denotes the sample autocovariance function of the time series x_t. $\hat{S}^{(p)}(f)$ is defined over the interval $[-f_{(n)}, f_{(n)}]$, where $f_{(n)}$ is called the *Nyquist frequency* and is given by

$$f_{(n)} = \frac{1}{2\Delta t}.$$

Although $\hat{S}^{(p)}(f)$ is an asymptotically unbiased estimator of the true spectral density function $S(f)$, it is well known that $\hat{S}^{(p)}(f)$ may be badly biased in some cases. There are two common techniques for reducing the bias in the periodogram: tapering and prewhitening. The latter will be described in Sect. 13.2.1.4.

Data tapering leads to the *direct spectral estimator*, which is defined by

$$\hat{S}^{(d)}(f) = \Delta t \left| \sum_{t=1}^{n} h_t x_t e^{-i2\pi f t \Delta t} \right|^2,$$

where $\{h_t, t = 1, \ldots, n\}$ is called the data taper sequence with $\sum_{t=1}^{n} h_t^2 = 1$.

Because $\hat{S}^{(d)}(f)$ is defined for all $f \in [-f_{(n)}, f_{(n)}]$, we can smooth it using a continuous convolution over a continuous set of frequencies. Thus, we consider an estimator of the form

$$\hat{S}^{(\mathrm{lw})}(f) = \int_{-f_{(n)}}^{f_{(n)}} W_m(f-\nu)\hat{S}^{(d)}(\nu)\,d\nu$$
$$= \Delta t \sum_{h=-(n-1)}^{n-1} w_{h,m}\hat{s}_h^{(d)} e^{-i2\pi f h \Delta t},$$

where $w_{h,m}$ and $W_m(\cdot)$ are a Fourier transform pair and $w_{h,m} = 0$ for $|h| \geq n$. $\hat{s}_h^{(d)}$ is the estimator of the autocovariance sequence corresponding to $\hat{S}^{(d)}(f)$, i.e., its inverse Fourier transform.

The function $W_m(\cdot)$ is a symmetric real-valued $2f_{(n)}$ periodic function for all choices of m, which is square integrable over $[-f_{(n)}, f_{(n)}]$, and m is a smoothing parameter that controls the degree of smoothing. $W_m(\cdot)$ is called a *smoothing window*, its inverse Fourier transform $w_{h,m}$ is called a *lag window*. Hence, we call $\hat{S}^{(\mathrm{lw})}(f)$ a *lag window spectral estimator* of $S(f)$.

Further details about the above described nonparametric spectral density estimators, their statistical properties and the different smoothing windows may be found in Priestley (1981) or Percival and Walden (1993).

13.2.1.3 Parametric Estimation

The most widely used form of parametric spectral density estimation uses an autoregressive model of order p as the underlying functional form for $S(f)$. A stationary AR(p) process $\{x_t, t \in \mathbb{Z}\}$ with zero mean satisfies the equation

$$x_t - \sum_{j=1}^{p} \phi_j x_{t-j} = \varepsilon_t,$$

where ε_t is a white noise process with zero mean and variance σ_ε^2. Thus, the spectral density function satisfies the following equation

$$\left| 1 - \sum_{j=1}^{p} \phi_j e^{-i2\pi f j \Delta t} \right|^2 S(f) = \Delta t \sigma_\varepsilon^2.$$

Substituting the maximum likelihood or least squares estimators of the model parameters, denoted by $\hat{\phi}_1, \ldots, \hat{\phi}_p$ and $\hat{\sigma}_\varepsilon^2$, we obtain a *parametric spectral density estimator*

$$\hat{S}^{(\mathrm{AR})}(f) = \frac{\Delta t \hat{\sigma}_\varepsilon^2}{|1 - \sum_{j=1}^{p} \hat{\phi}_j e^{-i2\pi f j \Delta t}|^2}, \quad |f| \leq f_{(n)}.$$

13.2.1.4 Semi-Parametric Estimation

If estimating the spectral density function the lag window spectral density estimator and the parametric spectral density estimator are the limiting versions prewhitening the process beforehand. The first approach leads to no prewhitening and the latter corresponds to total prewhitening if we assume that the process is a finite-order autoregressive process with known order p.

Let $\{y_t, t = 1, \ldots, n\}$ denote the observed values of a second-order stationary process with zero mean. Then the *prewhitened spectral density estimate* originally suggested by Blackman and Tukey (1958) is defined as

$$\hat{S}(f) = \frac{\hat{S}_r^{(\mathrm{lw})}(f)}{|1 - \sum_{j=1}^{p} \hat{\phi}_j e^{-i2\pi f j \Delta t}|^2}, \quad |f| \le f_{(n)}, \tag{13.3}$$

where $\hat{S}_r^{(\mathrm{lw})}(f)$ is a lag window spectral density estimate of the prediction residuals $r_t = y_t - \sum_{j=1}^{p} \hat{\phi}_j y_{t-j}$, $t = p+1, \ldots, n$.

13.2.2 Robust Spectral Density Estimation

13.2.2.1 Robust Prewhitening

Let $\{y_t, t = 1, \ldots, n\}$ again denote the observed process which is assumed to be second-order stationary and to have mean zero. The cleaning operator C maps the original data y_t into the cleaned data Cy_t. In the context of the AO model (13.1), we want the Cy_t to reconstruct the core process x_t, and so we will use the labeling $Cy_t = \hat{x}_{t|t}$, where $\hat{x}_{t|t}$ denotes an estimate of x_t at time t. The second index of $\hat{x}_{t|t}$ should indicate that the kind of data cleaning procedure we have in mind here is a robust filtering procedure which uses the past and present data values $y_1, \ldots, y_t$ to produce a cleaned filter estimate $\hat{x}_{t|t}$ of x_t, $t = 1, \ldots, n$. For AO models with a fraction of contamination γ not too large, it turns out that the data cleaner has the property that $Cy_t = y_t$ most of the time, that is about $(1-\gamma) \times 100$ percent of the time.

The filter-cleaner procedure involves a robust estimation of an autoregressive approximation to the core process x_t of order p, with estimated coefficients $\hat{\phi}_1, \ldots, \hat{\phi}_p$. Now, the residual process

$$r_t = Cy_t - \sum_{i=1}^{p} \hat{\phi}_i Cy_{t-i}, \quad t = p+1, \ldots, n, \tag{13.4}$$

can easily be formed. Since cleaned data are used to obtain these residuals, and the $\hat{\phi}_i$ are robust estimates, the transformation (13.4) is called a *robust prewhitening operation*. The benefit in the use of prewhitening in the context of spectral density estimation is to reduce the bias, i.e., the transfer of power from one frequency region of the spectral density function to another, known as leakage (cf. Blackman and Tukey 1958).

The robust spectral density estimate is then based on the robust prewhitening operation (13.4) and the prewhitened spectral density estimator (13.3) described above.

13.2.2.2 The Robust Filter–Cleaner Algorithm

The robust filter–cleaner algorithm as presented in the paper of Martin and Thomson (1982) is an approximate conditional-mean (ACM) type filter motivated by Masreliez's result (Masreliez 1975). It relies on the p-th order autoregressive approximation of the underlying process x_t, which can be represented in state-space form as follows. Assuming that x_t satisfies

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t$$

the state space model can be written as

$$\begin{aligned} \mathbf{x}_t &= \mathbf{\Phi}\mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t, \\ y_t &= x_t + v_t, \end{aligned}$$

with

$$\begin{aligned} \mathbf{x}_t &= (x_t, x_{t-1}, \ldots, x_{t-p+1})^\top, \\ \boldsymbol{\varepsilon}_t &= (\varepsilon_t, 0, \ldots, 0)^\top, \end{aligned}$$

and

$$\mathbf{\Phi} = \begin{pmatrix} \phi_1 & \cdots & \phi_{p-1} & \phi_p \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix}.$$

Additionally, we set

$$\operatorname{cov}(\boldsymbol{\varepsilon}_t) = \mathbf{Q} = \begin{pmatrix} \sigma_\varepsilon^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

and

$$\operatorname{var}(v_t) = \mathbf{R} = \sigma_0^2.$$

The algorithm computes robust estimates $\hat{\mathbf{x}}_{t|t}$ of the unobservable $\mathbf{x}_t$ according to the following recursion:

$$\hat{\mathbf{x}}_{t|t} = \mathbf{\Phi}\hat{\mathbf{x}}_{t-1|t-1} + \frac{\mathbf{m}_{.1,t}}{s_t^2} s_t \psi\left(\frac{y_t - \hat{y}_{t|t-1}}{s_t}\right)$$

with $\mathbf{m}_{.1,t}$ being the first column of $\mathbf{M}_t$ which is computed recursively as

$$\begin{aligned} \mathbf{M}_{t+1} &= \mathbf{\Phi}\mathbf{P}_t \Phi^\top + \mathbf{Q}, \\ \mathbf{P}_t &= \mathbf{M}_t - w\left(\frac{y_t - \hat{y}_{t|t-1}}{s_t}\right)\frac{\mathbf{m}_{.1,t}\mathbf{m}_{.1,t}^\top}{s_t^2}, \end{aligned}$$

where $\mathbf{M}_t$ and $\mathbf{P}_t$ are the prediction and filtering error-covariance matrices. The weight function w is defined by

$$w(r) = \frac{\psi(r)}{r},$$

where ψ stands for some robustifying psi-function. The scale s_t is set to

$$s_t^2 = m_{11,t}, \tag{13.5}$$

where $m_{11,t}$ is the first element of $\mathbf{m}_{.1,t}$. s_t^2 is an estimate of the variance of the observation prediction distribution conditioned on past observations. Further, $\hat{y}_{t|t-1}$ denotes a robust one-step-ahead prediction of y_t based on $\mathbf{Y}_{t-1} = \{y_1, \dots, y_{t-1}\}$, and is given by

$$\hat{y}_{t|t-1} = (\boldsymbol{\Phi}\hat{\mathbf{x}}_{t-1|t-1})_1.$$

Finally, the cleaned process at time t results in

$$\hat{x}_{t|t} = (\hat{\mathbf{x}}_{t|t})_1.$$

It should be noted that if ψ is the identity function and $w \equiv 1$, and (13.5) is replaced by $s_t^2 = m_{11,t} + \sigma_0^2$ with $\sigma_0^2 = \text{var}(v_t)$ in the AO model, the above recursions are those of the Kalman filter. The use of $\sigma_0^2 = 0$ in (13.5) corresponds to the assumptions that $v_t = 0$ a large fraction of time and that a contaminated normal distribution with degenerate central component (13.2) provides a reasonable model. The psi-function ψ and the weight function w which are essential to obtain robustness should be bounded and continuous. Additionally, it is highly desirable that both have zero values outside a bounded, symmetric interval around the origin. Furthermore, $\psi(s)$ is odd and should resemble the identity function for small values of s (see Martin 1979).

13.2.3 Small Simulation Study

Several methods to robustly estimate the spectral density function were compared by a simulation study. The compared methods are: robust Fourier transform using trimmed mean or Rousseeuw's minimum covariance determinant (MCD) estimator (Spangl 2008), methods based on a highly robust autocovariance function (Ma and Genton 2000) or on Spearman's rank correlation coefficient (Ahdesmäki et al. 2005), and robust prewhitening using an ACM type filter (Martin 1979) or the robust least squares (rLS) filter (Ruckdeschel 2001). The first two methods are robustified versions of the periodogram estimator, the next two are based on robust autocovariance functions, and the last two use robust filter algorithms to clean the process first and estimate the spectral density function afterwards. For further details about the different methods the reader is referred to Spangl (2008) and Spangl and Dutter (2007).

Table 13.1 Median L_2-distance (corresponding median absolute deviation given in parentheses)

Contamination	0 %	5 %	10 %	15 %	20 %
Robust Fourier Transform					
MCD	14.5 (0.992)	14.4 (1.021)	14.2 (1.167)	14.1 (1.203)	14.0 (1.246)
trimmed mean	11.1 (2.740)	11.0 (2.658)	10.9 (2.595)	10.9 (2.400)	11.4 (2.327)
Robust Autocovariance Function					
Ma & Genton	11.9 (1.756)	11.7 (1.617)	11.6 (1.596)	12.0 (1.288)	12.3 (1.239)
Spearman	11.5 (1.763)	10.8 (1.899)	10.5 (1.916)	10.7 (1.787)	11.2 (1.903)
Robust Filter Algorithms					
ACM	9.1 (4.644)	9.1 (4.726)	9.3 (4.827)	9.3 (4.386)	9.5 (4.120)
rLS	10.4 (5.097)	10.5 (5.410)	10.3 (5.011)	10.7 (4.709)	11.5 (3.552)

The outline of our simulation study is as follows: First a core process x_t of length $n = 100$ is simulated. x_t is chosen to be an autoregressive process of order 2 given by

$$x_t = x_{t-1} - 0.9x_{t-2} + \varepsilon_t,$$

with $\varepsilon_t \sim \mathcal{N}(0,1)$. Additionally, the additive outliers v_t are simulated from a contaminated normal distribution with degenerate central component (13.2) with $\sigma^2 = 10^2$. The contamination γ varies from 0 % to 20 % by steps of 5 %. That means that with probability γ, v_t is an additive outlier with $v_t \neq 0$. To obtain the contaminated process y_t, the v_t's are added to the core process x_t. For each level of contamination this is done 400 times.

For each contaminated series the spectral density function is robustly estimated using the above selected methods. Then, the deviation of each estimated spectral density function from the true spectral density function is measured in the sense of the L_2-norm, i.e.,

$$\mathrm{err}_{\hat{S}(f)} := \left\| \hat{S}(f) - S(f) \right\| = \left(\int \left(\hat{S}(f) - S(f) \right)^2 df \right)^{1/2},$$

where $\hat{S}(f)$ and $S(f)$ denote the estimated spectral density function and the true one.

The results are given in Table 13.1. For each method and each level of contamination the median L_2-distance together with its consistency-corrected median absolute deviation (MAD) is calculated which is given in parentheses. According to Table 13.1 the spectral density estimation based on the ACM type filter performs slightly better and will be further used to estimate the heart rate variability.

13.3 Application

13.3.1 Analysis of Heart Rate Variability

Heart rate variability (HRV) is composed of certain well-defined rhythms which contain information about the contribution of different regulatory mechanisms of cardiovascular control. In short-term HRV recordings, three spectral components can normally be distinguished: high frequency (HF, 0.15–0.4 Hz), low frequency (LF, 0.04–0.15 Hz), and very low frequency (VLF, 0–0.04 Hz) components. The HF component represents parasympathetic activity whereas the sympathetic nervous system is the main contributor of the LF component (cf. Hartikainen et al. 1998).

The analysis of heart rate variability as proposed in the review article by Pumprla et al. (2002) is, in fact, a combination of three short-term HRV recordings. The duration of each recording lasts 5 minutes as recommended by the Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology (1996). Moreover, the proposed modified orthostatic test where the individual lies supine for 5 minutes, stands for 5 minutes and lies supine again for another 5 minutes is also recommended when investigating patients with cardiovascular autonomic neuropathy in order to separate sympathetic from parasympathetic abnormalities.

As the tachogram recording is a discrete event series it is an irregularly time-sampled signal (cf. also Drews 1983). To obtain a regularly sampled series, we interpolate the original tachogram recording using cubic splines and resample at equidistant points in time. The resampling frequency has to be sufficiently high so that the Nyquist frequency of the spectral density function is not within the frequency range of interest. We therefore choose a resampling period of 0.25 seconds.

The outline of our dynamic Fourier analysis is as follows: a robust prewhitened spectral density estimate of the interpolated tachogram recording is calculated every 5 seconds for a time window with a duration of 256 seconds using the algorithm described above. For each window, the tachogram series is cleaned in a robust way by first using an ACM type filter. The hyperparameters of the approximating autoregressive process of order 5 are estimated robustly by bounded-influence autoregression (cf. Martin and Zeh 1978; Martin 1980). According to the order-selection rule described in Martin and Thomson (1982) an order of 5 seems to be sufficient. Then, the spectral density function is calculated using a prewhitened spectral density estimator. The lag window spectral estimate of the prediction residuals therein is obtained by using a $0th$-order discrete prolate spheroidal sequence (Thomson 1982) as data taper and a Parzen window (Parzen 1961) for smoothing. Finally, the results are displayed three dimensionally where we only plot the frequency range of interest, i.e., the LF and HF components. Here, it should be noted that in medicine the spectral densities are usually displayed on a metric scale and not on a logarithmic one.

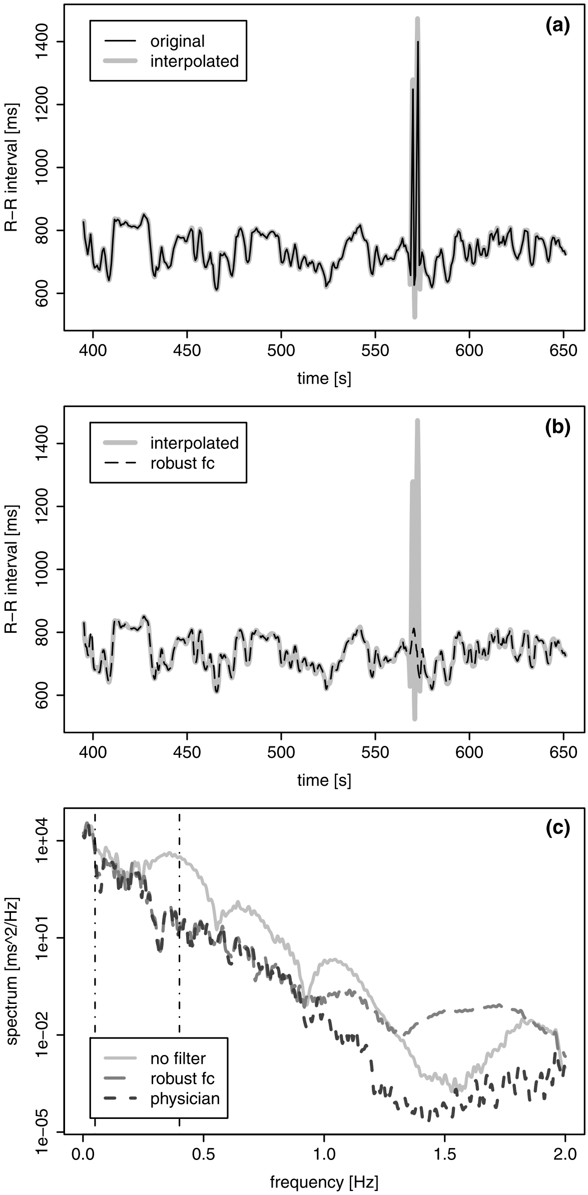


Fig. 13.3 Intermediate results of the suggested robust spectral analysis applied to the short-term HRV data displayed in Fig. 13.1

13.3.2 Results

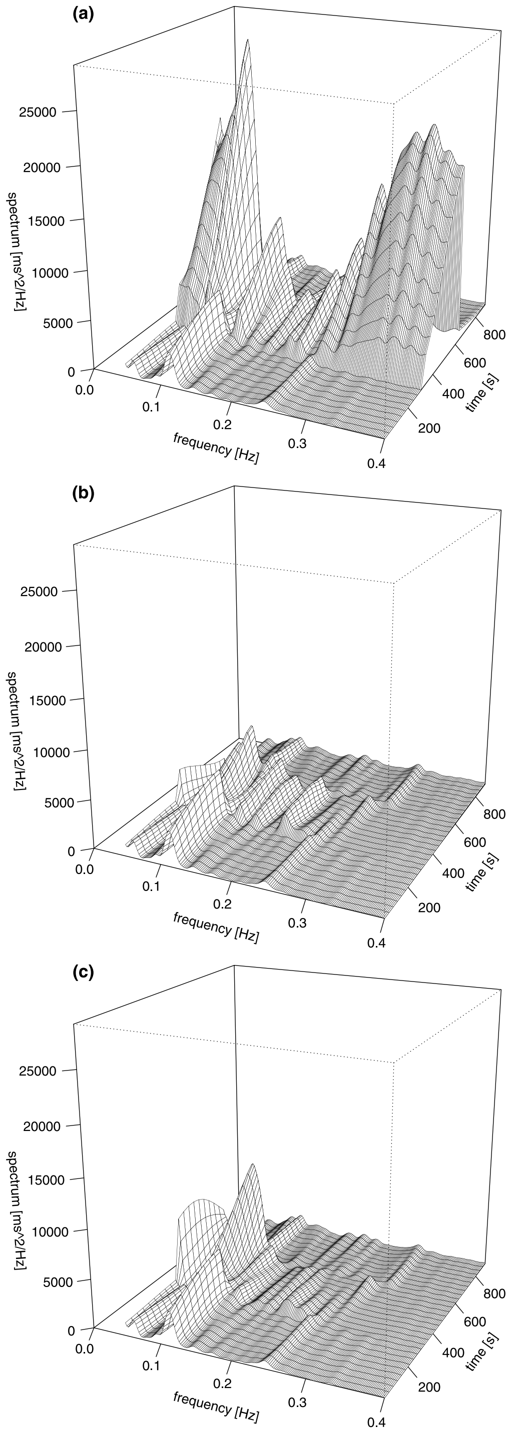
In Fig. 13.3, intermediate results of the proposed robust dynamic Fourier analysis are presented. To show how the suggested robust multi-step procedure works applied to the HRV data, we take one single time window of the data displayed in Fig. 13.1. The chosen example is the first 256-seconds window of the 5-minutes standing period indicated by the grey frame. Each of the three 5-minutes periods, that correspond to the supine position, standing, and the supine position again, is indicated in Fig. 13.1 by a vertical dashed line at the beginning and a dot-dashed one at the end.

Plot (a) of Fig. 13.3 shows the original tachogram recording (black line) along with the cubic spline interpolation (light grey line). As obviously seen the two are almost identical. In Plot (b), the interpolation result is displayed along with the robust filter estimate obtained by the approximate conditional-mean type filter. The filter estimate (dashed black line) is equivalent to the interpolated tachogram series (light grey line) in cases where no outliers are present. Additionally, it is not affected by outlying observations around 570 seconds that are caused by missed heart beats. Plot (c) shows several spectral density estimates of the HRV data. The prewhitened spectral density estimate of the robustly filtered tachogram series (long-dashed grey line) is similar in shape and power to the lag window spectral estimate of the tachogram series that was manually cleaned by the physician and will be considered as a benchmark in the following (dashed dark grey line). This correspondence is extremely well within the frequency range of interest, i.e., between 0.04 and 0.4 Hz, indicated by the two vertical dot-dashed lines. The difference between the two, especially visible for high frequencies, is negligible: first, it is not within the frequency range of interest and second, there is almost no difference in absolute values. The differences are only visible due to the logarithmic scale. Moreover, the lag window spectral estimate of the original tachogram series (light grey line) is markedly affected by outlying observations. We further note that, as we have chosen a resampling period of 0.25 seconds, the Nyquist frequency in this case is equal to 2 Hz.

The final result of the dynamic Fourier analysis is displayed in Fig. 13.4. Plot (a) shows the classical non-robust lag window spectral estimates of the original tachogram series. In Plot (b), the result of robust dynamic Fourier analysis is displayed whereas the result in Plot (c) is obtained by using the same estimator as in Plot (a) but now applied to the manually filter tachogram series. As before (cf. Fig. 13.3, Plot (c)), the results in Plot (b) and (c) of Fig. 13.4 are very similar. For the manually cleaned tachogram series (Plot (c)), the peaks in the estimated spectral density within the low frequency range, i.e., between 0.04 and 0.15 Hz, are slightly higher than for the robustly filtered series (Plot (b)). This is due to the fact that only single peaks are manually cleaned in the tachogram series whereas a whole area is smoothed when using a filter. Hence, the manually cleaned tachogram series is spikier as the filtered one which results in a higher spectral power for low frequencies. However, the spectral density estimates in Plot (a) are markedly affected in shape and power by outlying observations. Moreover, we note that Fig. 13.2 is equal to Plot (b), but for the latter we use the same scaling on the vertical axis as in Plot (a) to be able to compare them.

For further details and additional examples, the reader is referred to Spangl and Dutter (2012).

Fig. 13.4 Classical non-robust dynamic Fourier analysis (**a**), robust dynamic Fourier analysis (**b**), and classical Fourier analysis based on the manually filtered tachogram series (**c**) of the original short-term HRV data displayed in Fig. 13.1



13.4 Conclusions

As an application, we focused on the spectral analysis of short-term HRV data. To assess heart rate variability in the frequency domain, the spectral density function of the corresponding tachogram series has to be estimated.

Hence, to obtain a robust estimate of the spectral density function we suggest to use a multi-step procedure based on robust filtering. This algorithm first uses an ACM type filter (Martin 1979) based on an autoregressive approximation to eliminate outlying measurements, and then, a prewhitened spectral density estimator is applied. The presented method was compared to several other ones by simulation studies. These simulation experiments show that cleaning the series in a robust way first and calculating a prewhitened spectral density estimate afterwards leads to encouraging results. Moreover, the suggested procedure can be used to identify and mark outlying observations and, if slightly adapted, may also be used for online data processing.

The problem of estimating the hyperparameters was accomplished by bounded-influence autoregression. An alternative way would be to use a highly robust autocovariance function estimator (cf. Ma and Genton 2000) and calculate estimates of the hyperparameters via the Yule-Walker equations. Hyperparameters may also be obtained by computing a robust covariance matrix via the MCD algorithm (cf. Rousseeuw and Van Driessen 1999) and estimate the parameters again using the Yule-Walker equations. In Chap. 8 of their book (Maronna et al. 2006), propose to use τ-estimates instead.

We note that there are many other robust filter and data-cleaning procedures. In Spangl and Dutter (2005), the ACM type filter approach was compared with another approach proposed by Tatum and Hurvich (1993). This procedure also yields good results but tends to underestimate the core process slightly. Moreover, it is computationally intensive. Furthermore, a whole bundle of robust time series filters have already been implemented in R and are available in the R-package `robfilter` (Fried et al. 2012). Details may also be found in the contribution by Borowski, Fried and Imhoff, Chap. 12. However, these filters are specialized to reveal trends, trend changes or level shifts of an underlying, possibly nonstationary signal in the presence of outliers and smooth the underlying core process a lot. Hence, these filters are not applicable if the aim is the estimation of the spectral density function. A completely different approach is a robustified version of Welch's Overlapped Segment Averaging (WOSA) proposed by Chave et al. (1987). Although widely used in geophysical applications, it will only yield a good spectral density estimate if a small fraction of the data segments are contaminated by outliers.

Hence, up to date, the described multi-step procedure is the best method for estimating robustly the spectral density function, and it is well suited for accessing the heart rate variability in the frequency domain.

Acknowledgements All methods are implemented within the statistical programming language R (R Development Core Team 2009). On request the program code can be received from the authors. The rLS filter algorithm was implemented by Peter Ruckdeschel. Moreover, an R-package,

containing the BIAR algorithm, the ACM-type filter algorithm and the rLS filter algorithm, exists and is available at http://r-forge.r-project.org/projects/robkalman.

We would also like to thank J. Pumprla and K. Howorka, Department of Biomedical Engineering and Physics, Medical University of Vienna, for providing heart rate variability measurements and medical background information.

References

Ahdesmäki, M., Lähdesmäki, H., Pearson, R., Huttunen, H., & Yli-Harja, O. (2005). Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics*, *6*. doi:10.1186/1471-2105-6-117

Blackman, R., & Tukey, J. (1958). *The measurement of power spectra*. New York: Dover.

Chave, A., Thomson, D., & Ander, M. (1987). On the robust estimation of power spectra, coherences, and transfer functions. *Journal of Geophysical Research*, *92*, 633–648.

Denby, L., & Martin, R. (1979). Robust estimation of the first-order autoregressive parameter. *Journal of the American Statistical Association*, *74*, 140–146.

Drews, I. (1983). Zur statistischen Auswertung der Herzperiodendauer, unpublished.

Fox, A. (1972). Outliers in time series. *Journal of the Royal Statistical Society. Series B. Methodological*, *34*, 350–363.

Fried, R., Schettlinger, K., & Borowski, M. (2012). robfilter: robust time series filters. http://CRAN.R-project.org/package=robfilter.

Hartikainen, J., Tahvanainen, K., & Kuusela, T. (1998). Short-term measurement of heart rate variability. In M. Malik (Ed.), *Clinical guide to cardiac autonomic tests* (pp. 149–176). Dordrecht: Kluwer.

Howorka, K., Pumprla, J., Haber, P., Koller-Strametz, J., Mondrzyk, J., & Schabmann, A. (1997). Effects of physical training on heart rate variability in diabetic patients with various degrees of cardiovascular autonomic neuropathy. *Cardiovascular Research*, *34*, 206–214.

Howorka, K., Pumprla, J., & Schabmann, A. (1998). Optimal parameters of short-term heart rate spectrogram for routine evaluation of diabetic cardiovascular autonomic neuropathy. *Journal of the Autonomic Nervous System*, *69*, 164–172.

Jones, A., & Hollinger, K. (1997). Spectral analysis of the KTB sonic and density logs using robust nonparametic methods. *Journal of Geophysical Research*, *102*, 18391–18403.

Kleiner, R., Martin, R., & Thomson, D. (1979). Robust estimation of power spectra. *Journal of the Royal Statistical Society. Series B. Methodological*, *41*, 313–351.

Ma, Y., & Genton, M. (2000). Highly robust estimation of the autocovariance function. *Journal of Time Series Analysis*, *21*, 663–684.

Maronna, R., Martin, R., & Yohai, V. (2006). *Robust statistics: theory and methods*. New York: Wiley.

Martin, R. (1979). Approximate conditional-mean type smoothers and interpolators. In R. Gasser (Ed.), *Smoothing techniques for curve estimation*, New York: Springer.

Martin, R. (1980). Robust estimation of autoregressive models. In D. Brillinger & G. Tiao (Eds.), *Directions in time series* (pp. 228–254). Hayward: Institute of Mathematical Statistics.

Martin, R., & Thomson, D. (1982). Robust-resistant sprectrum estimation. In *Proceedings of the IEEE* (Vol. 70, pp. 1097–1115).

Martin, R., & Zeh, J. (1978). *Generalized M-estimates for autoregressions, including small-sample efficiency robustness*. Technical report 214, Deptartment of Electrical Engineering, University of Washington, Seattle.

Masreliez, C. (1975). Approximate non-Gaussian filtering with linear state and observation relations. In *IEEE transactions on automatic control* (pp. 107–110). AC-20.

Parzen, E. (1961). Mathematical considerations in the estimation of spectra. *Technometrics*, *3*, 167–190.

Percival, D., & Walden, A. (1993). *Spectral analysis for physical applications: multitaper and conventional univariate techniques*. Cambridge: Cambridge University Press.

Priestley, M. (1981). *Spectral analysis and time series: Vol. 1. Probability and mathematical statistics*. London: Academic Press.

Pumprla, J., Howorka, K., Groves, D., Chester, M., & Nolan, J. (2002). Functional assessment of heart rate variability: physiological basis and practical applications. *International Journal of Cardiology*, *84*, 1–14.

R Development Core Team (2009). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Rousseeuw, P., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*, 212–223.

Ruckdeschel, P. (2001). *Bayreuther Mathematische Schriften: Vol. 64. Ansätze zur Robustifizierung des Kalman-Filters*. Bayreuth: Universität Bayreuth.

Spangl, B. (2008). *On robust spectral density estimation*. Ph.D. Thesis, Department of Statistics and Probability Theory, Vienna University of Technology, Vienna.

Spangl, B., & Dutter, R. (2005). On robust estimation of power spectra. *Australian Journal of Statistics*, *34*, 199–210.

Spangl, B., & Dutter, R. (2007). Estimating spectral density functions robustly. *REVSTAT Statistical Journal*, *5*, 41–61.

Spangl, B., & Dutter, R. (2012). Analyzing short-term measurements of heart rate variability in the frequency domain using robustly estimated spectral density functions. *Computational Statistics & Data Analysis*, *56*, 1188–1199.

Task Force of The European Society of Cardiology, The North American Society of Pacing and Electrophysiology (1996). Heart rate variability. *Circulation*, *93*, 1043–1065. doi:10.1161/01.CIR.93.5.1043

Tatum, L., & Hurvich, C. (1993). A frequency domain approach to robust time series analysis. In S. Morgenthaler Ronchetti (Ed.), *New directions in statistical data analysis and robustness*, Basel: Birkhäuser.

Thomson, D. (1982). Spectrum estimation and harmonic analysis. In *Proceedings of the IEEE* (Vol. 70, pp. 1055–1096). New York: IEEE Press.

Thomson, D. (1994). An overview of multiple-window and quadratic-inverse spectrum estimation methods. In *Proceedings of the IEEE ICASSP* (Vol. 6, pp. 185–194). New York: IEEE Press.

Chapter 14
Robustness in Statistical Forecasting

Yuriy Kharin

14.1 Introduction

Many applied problems in engineering, economics, finance, medicine lead to an important problem of mathematical statistics—statistical forecasting of time series. According to the Webster encyclopedic dictionary (Maclaren 1946), "forecasting is an activity aimed at computing or predicting some future events or conditions based on rational analysis of the relevant data". The mathematical substance of the statistical forecasting problem is quite simple: estimate the future value $x_{T+\tau} \in \mathbb{R}^d$ of the d-variate time series $\tau \in \mathbb{N}$ steps ahead from $T \in \mathbb{N}$ successive observations $\{x_1, \ldots, x_T\} \subset \mathbb{R}^d$.

We can distinguish two stages in the history of attacks on the forecasting problem. The research on the *first stage* (before the year 1974) was oriented to the development of forecasting statistics that minimize the mean square forecast risk (error) for a set of simple mathematical models, e.g., stationary time series with some known spectral density, time series with a trend from some known parametric family, ARIMA time series; reviews of these classical results are given in Box and Jenkins (1976), Anderson (1971), Bowerman and O'Connel (1993).

In the seventieth of the last century, it was detected by many researchers that the true risk values of the "optimal" forecasting algorithms for the real statistical data are much more than the expected theoretical ones. In the lecture at the World Congress of Mathematicians in 1974, Peter Huber has explained the reason of this strange situation (Huber 1974): "Statistical inferences (including statistical forecasts) depend only in part upon the observations. An equally important base is formed by prior assumptions about the underlying situation". The system of prior assumptions is called the hypothetical data model M_0. In applied problems the assumptions of the hypothetical model are often distorted, and this fact leads to the

Y. Kharin (✉)
Research Institute for Applied Problems of Mathematics and Informatics, Belarusian State University, Independence av. 4, Minsk 220030, Belarus
e-mail: kharin@bsu.by

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_14, © Springer-Verlag Berlin Heidelberg 2013

instability of the "optimal" forecasting statistics that are optimal only under M_0. Huber has proposed to construct robust statistical inferences that are "weak-sensitive w.r.t. small distortions of the hypothetical model M_0". This event has opened the *second stage* in statistical forecasting of time series.

In last years, significant contributions to the theory of robust statistical data analysis were made by J. Tukey, P. Huber, F. Hampel, U. Gather, C. Becker, C. Croux, R. Dutter, R. Fried, P. Filzmoser, M. Genton, R. Maronna, R.D. Martin, S. Morgenthaler, H. Oja, D. Pena, H. Rieder, E. Ronchetti, P.J. Rousseeuw, V.J. Yohai, S. Van Aelst. The state of the research in this field is discussed in the nice analytic reviews prepared by Davies and Gather (2004, 2005), Gather et al. (2011), Maronna et al. (2006) and by Morgenthaler (2007). Professor Ursula Gather has organized a research network in robust statistical analysis for complex data structures that accumulate researchers from all over the world to attack the topical robustness problems. The majority of publications on robustness in statistical data analysis of time series are concentrated on estimation of parameters and hypotheses testing. Although these problems are fundamentals, they do not cover completely the problem of robustness in statistical forecasting of time series (see, e.g., Gelper et al. 2010). This chapter is devoted to topical problems of robust forecasting of time series. In Sect. 14.2, we give short mathematical description and classification of typical distortions for hypothetical models based on the reviews that were indicated and also on applications. Section 14.3 presents characteristics of robustness in forecasting. Sections 14.4 and 14.5 are devoted to robustness in statistical forecasting of time series under distorted regression and autoregression models respectively.

14.2 Distortions of Hypothetical Models for Time Series

Introduce the notation: $x_t \in \mathbb{R}^d$ is an observed d-variate time series with discrete time $t \in \mathbb{Z}$; $X = ({x'}_1, \dots, {x'}_T)' \subset \mathbb{R}^{Td}$ is the composed vector-column of observations for T time moments (prime symbol means transposition), $x_{T+\tau} \in \mathbb{R}^d$ is a non-observable random vector to be predicted at the future time moment $T + \tau$, $\tau \in \mathbb{N}$ (in economic applications the value T is called "the base of forecasting", τ is called "the horizon of forecasting"). The probability model of observed time series under distortions is determined by a family of probability measures

$$\left\{P^{\varepsilon}_{T,\theta^0}(A), A \in B^{Td} : T \in \mathbb{N}, \theta^0 \in \Theta \subseteq \mathbb{R}^m, \varepsilon \in [0, \varepsilon_+]\right\},$$

where B^{Td} is the Borel σ-algebra in $\mathbb{R}^{Td}$, θ^0 is an unknown true value of model parameters, ε is the distortion level, $\varepsilon_+ \geq 0$ is its maximal admissible value. If $\varepsilon_+ = 0$, then the distortions are absent, and we have the hypothetical model M_0.

A short scheme of classification for typical distortions of the hypothetical model M_0 is presented in Fig. 14.1 (a more detailed scheme of classification can be found in Kharin 2008). Let us give mathematical description of these distortions.

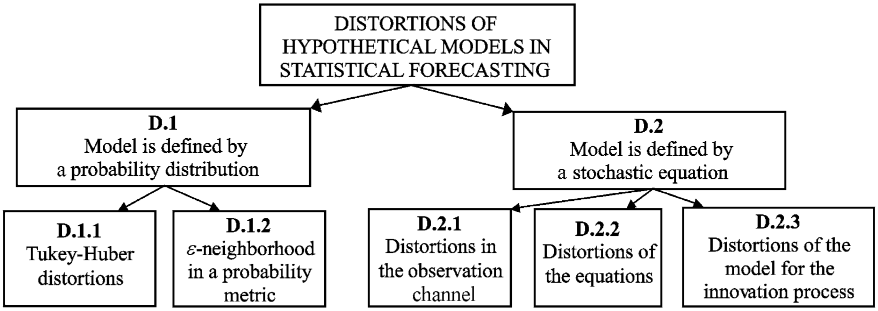


Fig. 14.1 Classification for types of distortions in statistical forecasting

With respect to the form of presentation of the hypothetical model M_0, the set of all types of distortions can be split into two classes: the model is in an explicit form (**D**.**1**), i.e. in the form of some hypothetical probability distribution $P^0(\cdot)$; the model is in an implicit form (**D**.**2**) determined by a stochastic equation: $x_t = G(x_{t-1},\dots,x_{t-s},u_t,u_{t-1},\dots,u_{t-L};\theta^0)$, $t \in \mathbb{Z}$, where $u_t \in \mathbb{R}^\nu$ is an innovation process (usually a white noise), $s, L \in \mathbb{N}$ are some natural numbers indicating the memory depth, $\theta^0 \in \Theta \subseteq \mathbb{R}^m$ is the vector of model parameters; $G(\cdot)$: $\mathbb{R}^{ds} \times \mathbb{R}^{\nu(L+1)} \times \Theta \to \mathbb{R}^d$ is some Borel function.

The Tukey–Huber distortions (**D**.**1**.**1**) for the observation vector X are described by a mixture: $p(X) = (1-\varepsilon)p^0(X) + \varepsilon h(X)$, where $p^0(\cdot)$ is some "non-distorted" (hypothetical) p.d.f., $h(\cdot)$ is a contaminating p.d.f., $p(\cdot)$ is the distorted p.d.f.

Distortions of the type **D**.**1**.**2** are described by ε-neighborhoods: $0 \le \rho(p(\cdot), p^0(\cdot)) \le \varepsilon$, where $\rho(\cdot)$ is some probability metric, e.g., the Kolmogorov metric, χ^2-metric (see Kharin and Shlyk 2009).

The class **D**.**2** consists of three subclasses. The subclass **D**.**2**.**1** describes distortions in the observation channel: instead of the clean data X^0 we observe a contaminated data set X, $X = H(X^0, V)$, where $X^0 = (x_k^0) \in \mathbb{R}^{Td}$ is a "non-observable history" of the process, $X \in \mathbb{R}^{Td}$ is the contaminated vector of observation results, that is the "observable history", $V = (v'_1,\dots,v'_T)' \in \mathbb{R}^{Tl}$ is a non-observable random vector of distortions (errors in the observation channel), $H(\cdot)$ is a function that describes the observation algorithm.

The subclass **D**.**2**.**1** includes six types of distortions. Additive (**D**.**2**.**1**.**1**) and multiplicative (**D**.**2**.**1**.**2**) distortions in the observation channel are described by the equations $x_t = x_t^0 + \varepsilon v_t$ and $x_t = (1+\varepsilon v_t)x_t^0$, $t \in \mathbb{Z}$, respectively, where $\{v_t\}$ are i.i.d. random variables, $\mathbb{E}\{v_t\} = 0, \mathbb{D}\{v_t\} = \sigma^2 < +\infty$. This type of distortions is considered in Fried and Gather (2002). The subclass **D**.**2**.**1**.**3** (ε-nonhomogeneities) includes the cases where the random vectors of distortions $\{v_t\}$ are non-identically distributed, but their probability distributions differ for not more than ε from the hypothetical one in some probability metric. The subclass **D**.**2**.**1**.**4** describes outliers in the data (see Gather et al. 2006 and also a review of Gather and Kale 1992). The replacement outliers (RO) and the additive outliers (AO) in the observation channel are described by the equations: $x_t = (1-\xi_t)x_t^0 + \xi_t v_t$ and

$x_t = x_t^0 + \xi_t v_t$, $t \in \mathbb{Z}$, respectively, where $\{\xi_t\}$ are i.i.d. Bernoulli random variables, $\mathbb{P}\{\xi_t = 1\} = 1 - \mathbb{P}\{\xi_t = 0\} = \varepsilon$, $\{v_t\}$ are random variables describing outliers, ε is the probability of an outlier appearance, and $\mathbb{E}\{v_t\}$, $\mathbb{D}\{v_t\}$ characterize the level of outliers (see also Kharin and Voloshko (2011) and the contribution by Galeano and Peña, Chap. 15). The subclass **D.2.1.5** considers missing values in $X^0 \in \mathbb{R}^{Td}$. Distortions generated by interval censoring of data X^0 are included into the subclass **D.2.1.6** (see Gather et al. 2011).

The subclass **D.2.2** describes distortions of the generating stochastic equation ("misspecification errors") and includes two types of distortions: parametric distortions (**D.2.2.1**), when instead of the true parameter value θ^0 we get (or estimate by statistical data) a different value $\tilde{\theta}$, with $|\tilde{\theta} - \theta^0| \le \varepsilon$, where ε is a distortion level; functional distortions (**D.2.2.2**), when instead of the true function $G(\cdot)$ we get data generated by a different function $\tilde{G}(\cdot)$, and in some metric $\|\tilde{G}(\cdot) - G(\cdot)\| \le \varepsilon$ (Kharin 2011a).

The subclass **D.2.3** describes distortions of the innovation process $u_t \in \mathbb{R}^{\nu}$, $t \in \mathbb{Z}$, in the generating stochastic equation: ε-nonhomogeneities (**D.2.3.1**), probabilistic dependence (**D.2.3.2**), "outliers" (**D.2.3.3**) (see the contribution by Spangl and Dutter, Chap. 13).

14.3 Robustness Characteristics in Forecasting

In statistical estimation and hypotheses testing, the most productive characteristics of robustness were introduced by Huber (1981) (the minimax approach), and by Hampel et al. (1986) (the approach based on the influence functions). For statistical forecasting problems, we use here (see also Kharin 2003, 2011b) the characteristics of robustness similar to the Γ-minimax risk criterion proposed by Berger (1985) and the robustness functionals in statistical pattern recognition considered in Kharin (1996).

Let $\hat{x}_{T+\tau} = f(X) : \mathbb{R}^{Td} \to R^d$ be any forecasting statistic. We evaluate its performance by the mean square risk of forecasting:

$$r_\varepsilon = r_\varepsilon(f) = \mathbb{E}_\varepsilon\bigl\{\|\hat{x}_{T+\tau} - x_{T+\tau}\|^2\bigr\} \ge 0, \quad \varepsilon \in [0, \varepsilon_+], \tag{14.1}$$

where $\mathbb{E}_\varepsilon\{\cdot\}$ is the expectation symbol w.r.t. the probability measure $P^{\varepsilon}_{T,\theta^0}(\cdot)$. If distortions are absent ($\varepsilon = 0$) and the hypothetical model M_0 is valid, the functional (14.1) is called the hypothetical risk $r_0 = r_0(f)$. Define the guaranteed (upper) risk:

$$r_+ = r_+(f) = \sup_{0 \le \varepsilon \le \varepsilon_+} r_\varepsilon(f), \tag{14.2}$$

where the supremum is taken w.r.t. all admissible distortions of M_0.

Further, let $\hat{x}^0_{T+\tau} = f^0(X; \theta^0)$ be an optimal forecasting statistic under the known hypothetical model M_0 that gives the minimal value to the hypothetical risk:

$$r_0 = r_0\bigl(f^0\bigr) = \inf_{f(\cdot)} r_0(f). \tag{14.3}$$

In practice, the family of the so-called “plug-in” forecasting statistics is often used: $\hat{x}_{T+\tau} = f(X) := f^0(X; \hat{\theta})$, where $\hat{\theta} \in \mathbb{R}^m$ is any consistent statistical estimator of the unknown parameter θ^0 based on the observed time series X.

For nonsingular cases, where $r_0 > 0$, define the risk instability coefficient κ as the relative increment of the guaranteed risk (14.2) w.r.t. the hypothetical risk (14.3):

$$\kappa = \kappa(f) = \bigl(r_+(f) - r_0\bigr)/r_0 \geq 0. \tag{14.4}$$

Another characteristic of robustness is the δ-admissible distortion level

$$\varepsilon^* = \varepsilon^*(\delta) = \sup\bigl\{\varepsilon \in [0, \varepsilon_+] : \kappa(f) \leq \delta\bigr\}, \quad \delta > 0. \tag{14.5}$$

It indicates the maximal level of distortions for which the relative increment of the risk is not greater than the fixed value $\delta \times 100\,\%$.

The smaller the value κ is and the greater the value ε^* is, the more robust the forecasting statistic is. The minimax robust forecasting statistic $\hat{x}^*_{T+\tau} = f^*(X)$ minimizes the risk instability coefficient (14.4):

$$\kappa\bigl(f^*\bigr) = \inf_{f(\cdot)} \kappa(f). \tag{14.6}$$

Also let us adjust the well known characteristic of “qualitative robustness” (see Hampel et al. 1986)—the Hampel breakdown point ε^{**}—to the forecasting problems: ε^{**} is the maximal fraction of “arbitrary large outliers” in time series X such that the considered forecasting statistic $f(X)$ is still bounded (see also the contribution by Müller, Chap. 5): $\varepsilon^{**} = \sup\{\varepsilon \in [0, 1] : \sup|f(X)| \leq C < +\infty\}$.

14.4 Robustness in Forecasting Under Distorted Regression Models

14.4.1 Robustness of the LS Forecasting Under Additive Outliers

Let the observed time series be described by a multiple linear regression model under additive outliers (type **D.2.1.4** of distortions):

$$x_t = \theta^{0'} \psi(z_t) + u_t + \xi_t v_t, \quad t \in \mathbb{N}, \tag{14.7}$$

where $z_t \in U \subseteq \mathbb{R}^M$ is a non-random observable vector of independent variables (regressors) at the time moment t; U is some “regressor space”; $\psi(\cdot) = (\psi_i(\cdot))$ is a vector-column of m linearly independent functions; $\{u_t\}$ are i.i.d. random variables with zero mean $\mathbb{E}\{u_t\} = 0$ and an unknown finite variance $\mathbb{D}\{u_t\} = \sigma^2$; $\theta^0 = (\theta^0_i) \in \mathbb{R}^m$ is a vector-column of m unknown regression coefficients; $\{\xi_t\}$ are i.i.d. Bernoulli random variables with

$$P\{\xi_t = 1\} = 1 - P\{\xi_t = 0\} = \varepsilon; \tag{14.8}$$

$\varepsilon \in [0, \varepsilon_+]$ is an unknown probability of outlier appearance; $\varepsilon_+ \in [0, \frac{1}{2})$ is some known upper bound for ε; the outliers $\{v_t\}$ are i.i.d. random variables with zero mean and an unknown variance

$$\mathbb{E}\{v_t\} = 0, \quad \mathbb{D}\{v_t\} = K \cdot \sigma^2, \quad K > 0; \tag{14.9}$$

$\{u_t\}$, $\{v_t\}$, $\{\xi_t\}$ are jointly independent.

It is well known, that the optimal forecasting statistic for the hypothetical model ($\varepsilon = 0$) under the known parameter value θ^0 is $\hat{x}^0_{T+\tau} = \theta^{0'} \psi(z_{T+\tau})$, and its hypothetical risk is $r_0 = \sigma^2$. Under an unknown value θ^0 the traditionally used forecasting statistic is the "plug-in" LS forecasting statistic:

$$\hat{x}_{T+\tau} = \hat{\theta}' \psi(z_{T+\tau}), \quad \hat{\theta} = \left(\Psi'_T \Psi_T\right)^{-1} \Psi'_T X, \quad \left|\Psi'_T \Psi_T\right| \neq 0, \tag{14.10}$$

where $\Psi_T = (\psi(z_1) \vdots \cdots \vdots \psi(z_T))'$ is a nonsingular $(T \times m)$-matrix; $\hat{\theta} = (\hat{\theta}_i)$ is the LS estimator of θ^0 consistent under the classical Eicker asymptotics (see Eicker 1963) for the minimal eigenvalue of the matrix $\Psi'_T \Psi_T$ at $T \to +\infty$: $\lambda_{\min}(\Psi'_T \Psi_T) \to +\infty$.

Introduce the notation:

$$(z)_+ = \max(z, 0), \quad C_T = \left(\Psi'_T \Psi_T\right)^{-1} \Psi'_T,$$
$$g(T, \tau) = \left(g_t(T, \tau)\right) = C'_T \psi(z_{T+\tau}) \in \mathbb{R}^T.$$

Robustness characteristics (defined in Sect. 14.3) for the LS forecasting statistic (14.10) are evaluated in Kharin (2011a).

Theorem 14.1 *If the observed time series x_t satisfies the regression model under additive outliers* (14.7)–(14.9), $T > m$, $|\Psi'_T \Psi_T| \neq 0$, *and the LS forecasting statistic* (14.10) *is used*, *then the instability coefficient equals*

$$\kappa_{\mathrm{LS}}(T, \tau) = \left\|g(T, \tau)\right\|^2 + \varepsilon_+ K\left(1 + \left\|g(T, \tau)\right\|^2\right), \tag{14.11}$$

and the δ-admissible distortion level is

$$\varepsilon^* = \varepsilon^*(\delta) = \min\left(\frac{1}{2}, \frac{(\delta - \|g(T, \tau)\|^2)_+}{K(1 + \|g(T, \tau)\|^2)}\right), \quad \delta > 0.$$

It is seen from (14.11) that the instability coefficient $\kappa(T, \tau)$ is the sum of two terms: the first term $\|g(T, \tau)\|^2$ is determined by prior uncertainty on θ^0 only, but the second one is generated by joint influence of outliers and prior uncertainty on θ^0.

14.4.2 Robustification by Huber Estimator

To robustify the LS forecasting statistic, let us use some robust estimator $\tilde{\theta} \in \mathbb{R}^m$ for the vector of regression coefficients θ^0 in (14.10) instead of the LS estimator $\hat{\theta}$:

$$\tilde{x}_{T+\tau} = \tilde{\theta}' \psi(z_{T+\tau}). \tag{14.12}$$

The Huber robust estimator (Huber 1981) $\tilde{\theta}$ belongs to the family of M-estimators, and it is defined by the following expressions:

$$\tilde{\theta} = \arg\min_{\theta} \sum_{t=1}^{T} \rho_H\big((x_t - \theta'\psi(z_t))/\tilde{\sigma}\big), \quad \rho_H(z) = \begin{cases} z^2/2, & z \in (-L, L), \\ Lz, & |z| \geq L, \end{cases}$$

where $\rho_H(\cdot)$ is the Huber loss function, $\tilde{\sigma} > 0$ is some robust estimator of the scale parameter, e.g., the estimator from Huber (1981), $L = L(\varepsilon)$ is the root of the equation: $2(\phi(L)/L - \Phi(-L)) = \varepsilon(1-\varepsilon)$, and $\phi(\cdot)$, $\Phi(\cdot)$ are the standard normal p.d.f. and cumulative distribution function respectively. This estimator $\tilde{\theta}$ minimizes the maximal value of the asymptotic mean square error for Gaussian probability distribution of $\{u_t\}$ under the **D.2.1.4** distortions.

By the asymptotic expansion method an approximation for the instability coefficient of the robust forecasting statistic (14.12) is constructed in Kharin (2008):

$$\kappa_H(T, \tau) \approx \big\|g(T, \tau)\big\|^2 + \varepsilon_+\big(K - 2L^2\big)\big\|g(T, \tau)\big\|^2. \tag{14.13}$$

From comparison of (14.11), (14.13) one can see that the instability coefficient for the robust forecasting statistic (14.12) is smaller than $\kappa_{\text{LS}}(T, S)$ for the value of the order $O(\varepsilon_+ L^2(\varepsilon_+)\|g(T, \tau)\|^2)$.

One disadvantage of the robust forecasting statistic (14.12) is its dependence on the usually unknown fraction ε of outliers in the observed data. Consider an approach that is free of this disadvantage.

14.4.3 Local-Median Robust Forecasting Statistic

Introduce the notation: $N_T = \{1, 2, \ldots, T\} \subset \mathbb{N}$ is the set of T observation times; $\Upsilon^{(l)} = \{t_1^{(l)}, \ldots, t_n^{(l)}\} \subseteq N_T$ is an l-th subset consisting of n different elements from N_T, $m \leq n \leq T, l = 1, \ldots, M$; M is the number of considered different subsets, $m \leq M \leq M_+ = \binom{T}{n}$; $\Psi_n^{(l)} = (\psi_j(z_{t_i^{(l)}}))$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, is the $(n \times m)$-submatrix of the matrix Ψ_T assumed to be non-singular, i.e. $\det(\Psi_n^{(l)'}\Psi_n^{(l)}) \neq 0$; $X_n^{(l)} = (x'_{t_1^{(l)}}, \ldots, x'_{t_n^{(l)}})' \in \mathbb{R}^n$ is the subsample of size n of the sample X.

Note that if the portion ε of outliers is small and $M = M_+$, then the portion of clean subsamples (containing no outliers) is $P_0 = \binom{[(1-\varepsilon)T]}{n}/\binom{T}{n}$ can be large enough ($[z]$ means the integer part of $z \in \mathbb{R}$). Using this fact, let us construct the l-th local LS estimator for θ^0 by the l-th subsample $X_n^{(l)}$: $\hat{\theta}^{(l)} = (\Psi_n^{(l)'}\Psi_n^{(l)})^{-1}\Psi_n^{(l)'}X_n^{(l)}$, and then the l-th local "plug-in" forecast $\hat{x}_{T+\tau}^{(l)} = \hat{\theta}^{(l)'}\psi(z_{T+\tau})$, $l = 1, \ldots, M$. The local-median (LM) forecasting statistic is proposed in Kharin (2008) as the sample median of M local forecasts $\{\hat{x}_{T+\tau}^{(l)}\}$:

$$\hat{x}_{T+\tau} = \text{med}\big\{\hat{x}_{T+\tau}^{(1)}, \ldots, \hat{x}_{T+\tau}^{(M)}\big\}. \tag{14.14}$$

Note that the subsample size n and the number of subsamples M are parameters of the LM method. If $n = T, M = 1$, then the LM forecast (14.14) is equivalent to the traditional LS forecast (14.10). If $M = M_+$, then the LM forecast uses all subsamples of size n from the observed sample X of size T.

The next theorem proved in Kharin (2011a) determines the Hampel breakdown point for the LM forecasting statistic (14.14).

Theorem 14.2 *Under Theorem 14.1 conditions, if $n < T$, $M = M_+$, then the Hampel breakdown point ε^{**} for the LM forecasting statistic (14.14) is the unique root ε in the segment $[0, 1 - n/T]$ of the algebraic equation of the order n:*

$$\prod_{t=0}^{n-1}\big(1 - \varepsilon - tT^{-1}\big) = (1-\alpha)\prod_{t=0}^{n-1}\big(1 - tT^{-1}\big), \quad \varepsilon \in [0, 1 - n/T], \tag{14.15}$$

with $\alpha = \lfloor (M_+ - 1)/2 \rfloor / M_+ = 1/2 + O(1/M_+)$.

Equation (14.15) can be solved numerically. There are explicit solutions of (14.15) for two cases: if $n = m = 1$, then $\varepsilon^{**} = \alpha$; if $n = 2$, then

$$\varepsilon^{**} = 1 - \frac{1}{2T} - \sqrt{(1-\alpha)\left(1 - \frac{1}{T}\right) + \frac{1}{4T^2}}.$$

Corollary 14.1 *If $n \geq m$ is fixed, but $T \to +\infty$, then $\varepsilon^{**} \to 1 - 2^{-1/n}$, and the optimal size of subsamples that maximizes the asymptotical Hampel breakdown point equals to the number of unknown regression coefficients: $n^* = m$.*

In Kharin (2008), an approximation of the risk instability coefficient for the LM forecasting statistic (14.14) is found under the asymptotics $T \to +\infty$, $M = M(T) \to +\infty$:

$$\kappa_\varepsilon(T, \tau) \approx \varepsilon K + \frac{\pi}{2M}\left(\sum_{r=0}^{n} \frac{(1-\varepsilon)^r \varepsilon^{n-r}\binom{n}{r}}{\sqrt{G^{(n,1)} + KG^{(n,r+1)}}}\right)^{-2},$$

where $G^{(n,r)} = \sum_{k=r}^{n}(g_k)^2$, $g = (g_k) = \Psi_n^{(l)}(\Psi_n^{(l)'}\Psi_n^{(l)})^{-1}\psi(z_{T+\tau})$, $G^{(n,n+1)} := 0$. The gains in the risk for the LM forecast w.r.t. the LS forecast (14.10) and the Huber forecast (14.12) can be evaluated using this approximation together with (14.11), (14.13).

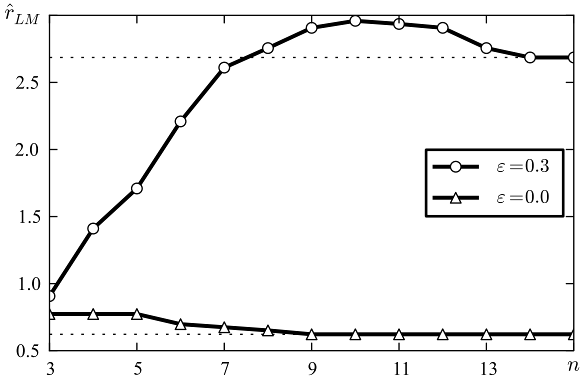


Fig. 14.2 Dependence of risk on subsample size n

Monte-Carlo simulations for the model (14.7) were made for the traditional LS forecasting statistic (14.10) and for the LM robust forecasting statistic (14.14) under the following conditions:

$$M=1,\quad z_t=t,\quad m=3,\quad \psi(z_t)=\big(1,t,t^2\big)',\quad \theta=(1,0.1,0.01)',$$
$$u_t\sim\mathcal{N}\big(0,\sigma^2\big),\quad v_t\sim\mathcal{N}\big(0,K\sigma^2\big),\quad \sigma^2=0.09,\quad K=50,\quad T=15,$$
$$\tau\in\{1,2,3,4,5\}.$$

Figure 14.2 illustrates dependence of the sample forecast risk $\hat{r}_{\mathrm{LM}}$ for the LM forecasting statistic (14.14) with $\tau=4$ (by the series of 10 Monte-Carlo simulations) on the size of local subsamples $n\in\{3,4,\ldots,15\}$ for two distortion levels: $\varepsilon=0$ (no outliers) and $\varepsilon=0.3$ (on the average, 30 % of outliers are present in observed time series). The results for $\varepsilon=0$ are shown as triangles, and for $\varepsilon=0.3$ as circles; dashed lines in Fig. 14.2 correspond to the risk of the LS forecasting algorithm (14.10) for $\varepsilon=0$ and $\varepsilon=0.3$. It is seen that under outliers the lowest LM forecast risk is attained at $n^*=m=3$.

Let us note that the proposed robust LM forecasting statistic (14.14) can be generalized for the case of d-variate regression time series $x_t\in\mathbb{R}^d$ using the multivariate Oja median (Oja 1983).

Let us mention one more approach to robust forecasting under distorted regression models of time series proposed by Gelper et al. (2010) based on smoothing of time series.

14.4.4 Nonparametric Distortions of Regression Functions

Let the observed time series be described by a multiple linear regression model under nonparametric distortions (the type **D.2.2.2** of distortions):

$$x_t=\theta^{0'}\psi(z_t)+\lambda(z_t)+u_t,\quad t\in\mathbb{N},\tag{14.16}$$

where $\lambda(\cdot)$: $\mathbb{R}^M \to \mathbb{R}^1$ is an unknown function that describes functional distortions; the other variables are defined as in (14.7). The regression function for the model (14.16) under $z_t = z$ is $\mu(z) = \mathbb{E}\{x_t|z\} = \mu_0(z) + \lambda(z)$, where $\mu_0(z) = \theta^{0'}\psi(z)$ is the hypothetical (assumed) regression function, but $\lambda(z) = \mu(z) - \mu_0(z)$ is some functional distortion that could be generated by the prior uncertainty, the complexity of the real observed process, or by special features of the observation channel.

Define three special types of functional distortions $\lambda(\cdot)$.

FD-1 Interval distortions: $\varepsilon_-(z) \le \lambda(z) \le \varepsilon_+(z)$, $z \in U \subseteq \mathbb{R}^M$, where $\varepsilon_\pm(z)$ are some boundary functions, in particular,

$$-\varepsilon \le \lambda(z) \le +\varepsilon, \quad z \in U; \tag{14.17}$$

here $\varepsilon \ge 0$ is the distortion level in the C-metric.

FD-2 Relative distortions: $|\lambda(z)|/|\theta^{0'}\psi(z)| \le \varepsilon$, $z \in U$, $\varepsilon \ge 0$.

FD-3 Distortions in the l_p-metric: $(\sum_{t=1}^T |\lambda(z_t)|^p + |\lambda(z_{T+\tau})|^p)^{\frac{1}{p}} \le \varepsilon$.

Theorem 14.3 *If the observed time series x_t satisfies the regression model* (14.16) *under interval distortions* **FD-1** *defined by* (14.17), *then the risk instability coefficient and the δ-admissible distortion level, respectively, are*

$$\kappa(T,\tau) = \left\|g(T,\tau)\right\|^2 + (\varepsilon/\sigma)^2\left(1 + \sum_{t=1}^T \left|g_t(T,\tau)\right|\right)^2,$$

$$\varepsilon_+(\delta) = \sigma\left(\left(\delta - \left\|g(T,\tau)\right\|^2\right)_+\right)^{1/2} \Big/ \left(1 + \sum_{t=1}^T \left|g_t(T,\tau)\right|\right).$$

It follows from Theorem 14.3 that if the level ε of the interval distortions (14.17) is higher than the critical value $\varepsilon_+(\delta)$, then the forecast risk r can exceed the value $r = (1+\delta)\cdot r_0$ that is $\delta \times 100$ % more than the hypothetical value $r_0 = \sigma^2$.

Similar results are given in Kharin (2011a) for the distortions **FD-2, FD-3**.

To construct a robust forecasting statistic under the distortion type **FD-1** consider a family of "plug-in" forecasting statistics based on M-estimators $\hat{\theta}$ (see, e.g., Huber 1981; Hampel et al. 1986):

$$\hat{x}_{T+\tau} = \hat{\theta}'\psi(z_{T+\tau}), \quad \hat{\theta} = \arg\min_{\theta} \sum_{t=1}^T \rho\big(x_t - \theta'\psi(z_t)\big), \tag{14.18}$$

where $\rho(z)$ is a convex, even, twice differentiable (almost everywhere on $\mathbb{R}$) loss function; a special case in the family (14.18) is the LS forecasting statistic with the quadratic loss function: $\rho(z) = 0.5z^2$. The concrete form of the loss function $\rho(\cdot)$ needs to be determined by the type of functional distortions. To compute $\hat{\theta}$ in (14.18), we have the following iterative procedure for the convex minimization

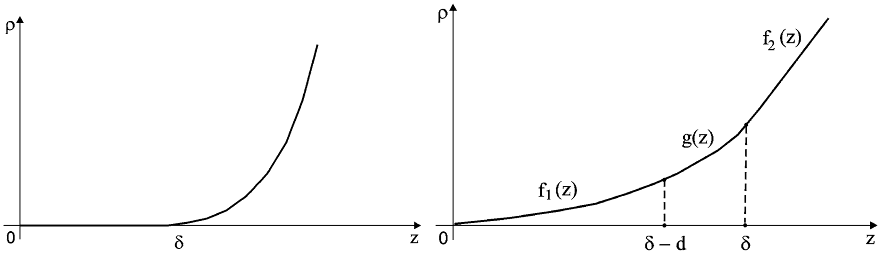


Fig. 14.3 The loss functions: *left*—for (14.20), *right*—for (14.21)

problem:

$$
\begin{aligned}
&\mu(z) = \rho'(z), \quad \nu(z) = \rho''(z);\\
&\theta_{(n+1)} = \theta_{(n)} + \big(D(\theta_n)\big)^{-1} M(\theta_{(n)}), \quad n = 1, 2, \ldots;\\
&D(\theta) = \Psi_T' \nu(\theta) \Psi_T, \quad M(\theta) = \Psi_T' \mu(\theta),\\
&\mu(\theta) = \big(\mu_t(\theta)\big) = \big(\mu\big(x_t - \theta'\psi(z_t)\big)\big),\\
&\nu(\theta) = \big(\nu_{tt'}(\theta)\big) = \operatorname{diag}\big(\nu\big(x_t - \theta'\psi(z_t)\big)\big), \quad t, t' = 1, \ldots, T.
\end{aligned}
\tag{14.19}
$$

Taking into account the distortion type **FD-1** defined by (14.17), let us construct the loss function $\rho(\cdot)$ in the special form (see Fig. 14.3):

$$
\rho(z) = 0.5 \cdot \mathrm{I}\big(|z| - \delta_\varepsilon\big)\big(z - \delta_\varepsilon \operatorname{sign}(z)\big)^2, \quad z \geq 0,
\tag{14.20}
$$

where $\delta_\varepsilon \geq 0$ is some tuning parameter of the algorithm, $\mathrm{I}(z) = \{1, z > 0; 0, z \leq 0\}$; if $\delta_\varepsilon = 0$, we get the LS forecasting statistic (14.10). Using (14.19), (14.20) we have: $\mu(z) = \mathrm{I}(|z| - \delta_\varepsilon)(z - \delta_\varepsilon \operatorname{sign}(z))$, $\nu(z) = \mathrm{I}(|z| - \delta_\varepsilon)$, $z \neq \pm\, \delta_\varepsilon$. The loss function (14.20) has a nonempty set with zero sensitivity of the objective function in (14.18), so the convergence properties of the iterative procedure (14.19) depend significantly on the starting point $\theta_{(1)}$. To avoid this defect, introduce a smoothed version of the loss function (14.20) (see Fig. 14.3):

$$
\rho(z) = f_1(z)\mathrm{I}_{[0, \delta_\varepsilon - d]}(z) + g(z)\mathrm{I}_{(\delta_\varepsilon - d, \delta_\varepsilon)}(z) + f_2(z)\mathrm{I}_{[\delta_\varepsilon, +\infty)}(z),
$$

where $f_1(z) = az^2$ is some parabola, $g(z)$ is a "connecting function", $f_2(z) = b(z - \alpha)^2 + \beta$ is some other parabola, d is the length of the "connecting zone", $a, b > 0$. Choosing $a \ll b$ we achieve a small but not-zero sensitivity of the objective function in (14.18). Using a cubic "connecting function" $g(z)$ and the assumption of twice differentiability of $\rho(z)$ we get the "smoothed" loss function ($z \geq 0$):

$$
\rho(z) = \begin{cases}
az^2, & 0 \leq z \leq \delta_\varepsilon - d,\\
az^2 + \frac{b-a}{3d}(z + d - \delta_\varepsilon)^3, & \delta_\varepsilon - d < z < \delta_\varepsilon,\\
b\big(z - \frac{(b-a)(2\delta_\varepsilon - d)}{2b}\big)^2 + \frac{(b-a)(12a\delta_\varepsilon(\delta_\varepsilon - d) + d^2(b+3a))}{12b}, & z \geq \delta_\varepsilon.
\end{cases}
\tag{14.21}
$$

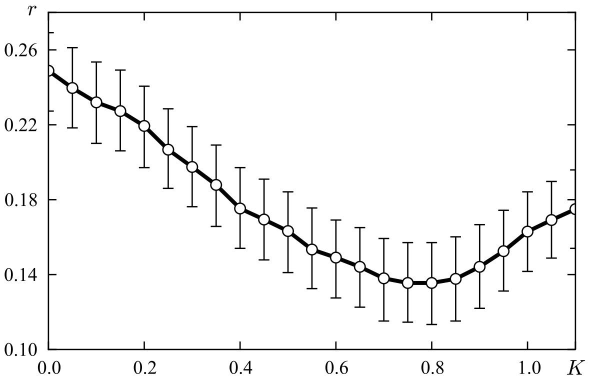


Fig. 14.4 Dependence of risk on coefficient $K = \delta/\varepsilon$

Note that if the errors $\{u_t\}$ have a Gaussian distribution, $\lambda(z_t) = \pm\varepsilon$, and $\lim_{T\to+\infty} \frac{1}{T}\sum_{t=1}^{T} \psi(z_t) = \bar{\psi}$, $\lim_{T\to+\infty} \frac{1}{T}\sum_{t=1}^{T} \psi(z_t)\psi'(z_t) = H_0$, then for $T \to +\infty$, the guaranteed risk has the asymptotic expression presented in Kharin (2011a):

$$\begin{aligned} r_+(T,\tau) \to \sigma^2 + \varepsilon^2 &+ 2\varepsilon\sigma G(\tilde{\delta}_\varepsilon, \tilde{\varepsilon})\left|\psi'(z)H_0^{-1}\bar{\psi}\right| \\ &+ \sigma^2 G^2(\tilde{\delta}_\varepsilon, \tilde{\varepsilon})\left(\psi'(z)H_0^{-1}\bar{\psi}\right)^2, \end{aligned} \tag{14.22}$$

where $\tilde{\delta_\varepsilon} = \delta_\varepsilon/\sigma$, $\tilde{\varepsilon} = \varepsilon/\sigma$,

$$G(x,y) = \frac{(x+y)\Phi(-x-y) + (-x+y)\Phi(-x+y) + \phi(x-y) - \phi(x+y)}{\Phi(-x-y) + \Phi(-x+y)}. \tag{14.23}$$

It follows from (14.22), (14.23), that the maximal gain of the robust forecasting statistic (14.18) w.r.t. to the LS forecasting statistic (14.10) is attained for the situations with $\varepsilon/\sigma \gg 1$, i.e., for the situations where the errors determined by functional distortions are much larger than the random observation errors. In Kharin (2011a), an approximation for the optimal value $\delta_\varepsilon^* \approx \varepsilon$ minimizing the risk instability coefficient is found.

Monte-Carlo experiments for the model (14.16) were made for the traditional LS forecasting statistic (14.10) and for the robust forecasting statistic (14.18) with the loss function (14.20) under following conditions:

$$\begin{aligned} &z_t = t, \quad \mu_0(z_t) = 3 - 0.5t + 0.05t^2, \quad \lambda(z_t) = \varepsilon\cos t, \quad m = 3, \\ &u_t \sim \mathcal{N}\left(0,\sigma^2\right), \quad \sigma^2 = 0.01, \quad T = 20, \quad \tau = 1, \quad a = 0.01, \quad b = 1, \quad \delta_\varepsilon = K\varepsilon, \end{aligned}$$

where $K = 0; 0.05; 0.10; \ldots; 1.50$. Point and 95 % confidence interval estimates for the risk of the forecasting statistic (14.18) are evaluated by 10^3 Monte-Carlo replications and are plotted in Fig. 14.4.

Note that the case $K = 0$ corresponds to the traditional LS forecasting statistic. It is seen from Fig. 14.4 that the robust forecasting statistic (14.18) with the optimal value $\delta_\varepsilon^* \approx \varepsilon$ significantly decreases the risk under distortions: $r_* = 0.13$ in comparison with the risk of the traditional LS forecasting statistic $r = 0.24$.

Let us mention that the robust forecasting statistic (14.18) can also be applied to multivariate time series by the following modification:

$$\hat{x}_{T+\tau} = \hat{\theta}\psi(z_{T+\tau}), \quad \hat{\theta} = \arg\min_{\theta} \sum_{t=1}^{T} \rho\bigl(\|x_t - \theta\psi(z_t)\|\bigr),$$

where $\hat{\theta} = (\hat{\theta}_{jk})$ is the estimator of the $(d \times m)$-matrix of regression coefficients.

14.5 Robustness in Forecasting Under Distorted Autoregression Models

14.5.1 Misspecification of Autoregression Coefficients

Consider now one more hypothetical model of observed data used in applied problems of statistical forecasting—the autoregression model AR(p) of order $p \in \mathbb{N}$:

$$x_t = \theta_1^0 x_{t-1} + \cdots + \theta_p^0 x_{t-p} + u_t, \quad t \in \mathbb{Z}, \tag{14.24}$$

where $\theta^0 = (\theta_1^0, \ldots, \theta_p^0)' \in \mathbb{R}^p$ is an unknown vector-column of p autoregression coefficients satisfying the stationarity condition (Anderson 1971), u_t is the innovation process assumed to be a sequence of i.i.d. Gaussian random variables with zero mean and finite unknown variance $\mathbb{E}\{x_t^2\} = \sigma^2$. The optimal τ-step-ahead forecasting statistic is linear (see, e.g., Anderson 1971):

$$\hat{x}^*_{T+\tau} = \sum_{j=1}^{p} \bigl(A_0^\tau\bigr)_{1j} x_{T+1-j}, \quad \tau \in \mathbb{N}, \tag{14.25}$$

where the $(p \times p)$-matrix A_0 is the companion matrix of the stochastic difference equation (14.24) in the block form:

$$A_0 = \begin{pmatrix} & \theta^{0'} & \\ \cdots & \cdots & \cdots \\ \mathrm{I}_{p-1} & \vdots & \mathbf{O}_{p-1} \end{pmatrix}, \tag{14.26}$$

I_p is the identity matrix of order p, $\mathbf{O}_{p-1}$ is the zero vector-column of size $p-1$. The minimal admissible risk for the forecasting statistic (14.25) is

$$r_0(\tau) = \sigma^2 \left(1 + \sum_{k=1}^{\tau-1} \bigl(\bigl(A_0^k\bigr)_{11}\bigr)^2\right).$$

Due to prior uncertainty, the true autoregression coefficient $\theta^0 \in \mathbb{R}^p$ is available with some specification error $\Delta\theta \in \mathbb{R}^p$, and thus the forecasting statistic is constructed from distorted values (**D.2.2.1** type of distortion): $\theta = \theta^0 + \Delta\theta$. The plug-in forecasting statistic uses the vector θ and is similar to (14.25), (14.26):

$$\hat{x}_{T+\tau} = \sum_{j=1}^{p} \left(A^{\tau}\right)_{1j} x_{T+1-j}, \qquad A = \begin{pmatrix} & \theta' & \\ \cdots & \cdots & \cdots \\ \mathbf{I}_{p-1} & \vdots & \mathbf{O}_{p-1} \end{pmatrix}. \tag{14.27}$$

Introduce the notation: $\lambda_{\max}(C)$ is the maximal eigenvalue of the matrix C,

$$B_{\tau} = \sum_{k=1}^{\tau-1} \left(A_0^k\right)_{11} \left(A_0^{\tau-k}\right)', \qquad \Sigma_p = \begin{pmatrix} \sigma_0 & \sigma_1 & \dots & \sigma_{p-1} \\ \sigma_1 & \sigma_0 & \dots & \sigma_{p-2} \\ \vdots & \vdots & \dots & \vdots \\ \sigma_{p-1} & \sigma_{p-2} & \dots & \sigma_0 \end{pmatrix}.$$

In the asymptotics of small specification errors $\varepsilon = \max|\Delta\theta| \to 0$ the instability coefficient for the forecasting statistic (14.27) satisfies the asymptotic expansion (presented in Kharin 2008): $\kappa(\tau) = \varepsilon^2 \lambda_{\max}(B_{\tau} \Sigma_p B_{\tau}{}')/r_0(\tau) + O(\varepsilon^3)$.

14.5.2 Distortions of Innovation Process

Let us discuss the situation, commonly occurring in practice, that the hypothetical AR(p) model (14.24) is affected by nonhomogeneities in the mean value of the innovation process (type **D.2.3.1** of distortions):

$$x_t = \theta_1^0 x_{t-1} + \dots + \theta_p^0 x_{t-p} + \tilde{u}_t, \quad \tilde{u}_t = \mu_t + u_t,\ t \in \mathbb{Z}, \tag{14.28}$$

where u_t is determined in (14.24), $\tilde{u}_t$ is the distorted innovation process, $\mu_t \in \mathbb{R}$ is some unknown deterministic function defining the distortion (see Kharin 2008, 2011b).

Theorem 14.4 *Under a specification error $\Delta\theta$ and nonhomogeneities (14.28) in mean values of the innovation process satisfying the inequalities*

$$T^{-1} \sum_{t=1}^{T} \mu_t^2 \le \varepsilon_1^2, \qquad \tau^{-1} \sum_{t=T+1}^{T+\tau} \mu_t^2 \le \varepsilon_2^2,$$

the risk instability coefficient of the τ-step-ahead autoregression forecasting statistic (14.28) based on T observations equals

$$\kappa(\tau) = \frac{1}{r_0(\tau)} \left(\sigma^2 \sum_{k=0}^{T-1} a^2(k,\tau) + \left(\varepsilon_1 \sqrt{\frac{1}{T} \sum_{k=0}^{T-1} a^2(k,\tau)} + \varepsilon_2 \sqrt{\frac{1}{\tau} \sum_{k=0}^{\tau-1} \left((A_0^k)_{11} \right)^2} \right)^2 \right)$$
$$\ge 0,$$

where $a(k,\tau) = ((A^{\tau} - A_0^{\tau}) A_0^k)_{11} \in \mathbb{R}$.

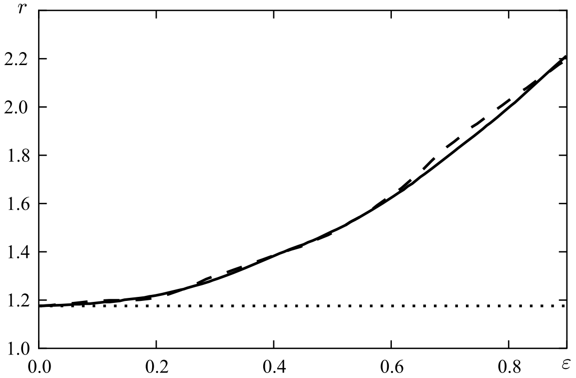


Fig. 14.5 Dependence of risk on distortion level ε

Consider now the situation where the innovation process is influenced by heteroscedasticity, i.e., functional distortion μ_t of the variance is present:

$$x_t = \theta_1^0 x_{t-1} + \cdots + \theta_p^0 x_{t-p} + \tilde{u}_t, \quad \tilde{u}_t = \mu_t u_t, \ t \in \mathbb{Z}. \tag{14.29}$$

Theorem 14.5 *For the* AR(p) *model under heteroscedasticity* (14.29) *and a specification error* $\Delta\theta$ *in the asymptotics* $\varepsilon = \max|\Delta\theta| \to 0$ *the risk instability coefficient of the forecasting statistic* (14.27) *satisfies the asymptotic expansion*:

$$\kappa(T,\tau) = \varepsilon^2 \lambda_{\max}\big(\beta(T,\tau)\big) + O\big(\varepsilon^3\big),$$

where $(A_0^t)_{\cdot 1}$ *is the first column of the matrix* A_0^t, $\beta(T,\tau) = (\sum_{i=0}^{\tau-1}(A_0^i)_{11} A_0^{\tau-i-1}) \times S(\sum_{i=0}^{\tau-1}(A_0^i)_{11} A_0^{\tau-i-1})'$, $S = \sigma^2 \sum_{t=0}^{+\infty}(A_0^t)_{\cdot 1}((A_0^t)_{\cdot 1})' \mu_{T-t}^2$.

Monte-Carlo simulations agree with the obtained theoretical results. Figure 14.5 presents results of computer simulations based on the model (14.28) with $p = 2$, $\theta^0 = (0.3, 0.4)'$, $\theta = (0.4, 0.5)'$, $\sigma^2 = 1$, $\mu_t = \varepsilon \cdot \sin(t)$, $\varepsilon_{(1)} = \varepsilon_{(2)} = \varepsilon$, $T = 40$, $\tau = 2$; 10^4 Monte-Carlo simulation rounds were performed. The solid line shows the theoretical dependence of risk $r_\varepsilon(\tau) = r_0(\tau)(1 + \kappa(\tau))$ (computed by Theorem 14.4) on distortion level ε; the dashed line represents experimental values of the risk, and the dotted line indicates the minimal risk $r_{\min} = r_0(\tau) + \sigma^2 \sum_{k=0}^{T-1} a^2(k,\tau)$ computed for $\mu_t \equiv 0$.

14.5.3 Bilinear Distortions of AR(p)

Most of the real-world processes have nonlinear nature. However, they are often analyzed by using simpler and better-studied linear models. In that case the key question is the quantitative effect of nonlinearity on the inferences based on a linear model. Here we study a nonlinear modification of the linear autoregression model—the bilinear model (see Kharin 2008).

Let the observed time series x_t have distortions of type **D.2.2** and satisfy the bilinear model BL$(p, 0, 1, 1)$ considered in Fan and Yao (2003):

$$x_t = \theta_1^0 x_{t-1} + \cdots + \theta_p^0 x_{t-p} + \varepsilon x_{t-1} u_{t-1} + u_t, \quad t \in \mathbb{Z}, \tag{14.30}$$

where ε is the bilinearity coefficient. If $\varepsilon = 0$, then the model (14.30) coincides with the hypothetical AR(p) model defined by (14.24). Note that the stationarity condition is (see Fan and Yao 2003): $\rho(A_0 \otimes A_0 + \sigma^2 \varepsilon^2 C \otimes C) < 1$, where $\rho(A)$ is the spectral radius of the matrix A, $\otimes$ is the Kronecker matrix product, and the $(p \times p)$-matrix C has zero elements except for $(C)_{1,1} = 1$.

Introduce matrices: $W = \mathrm{I}_{p+1} - W_1 - W_2$, $\bar{S} = (\bar{s}_{ij}) = S^{-1}$, $\bar{W} = (\bar{w}_{ij}) = W^{-1}$,

$$W_1 = \begin{pmatrix} 0 & \theta_1^0 & \theta_2^0 & \dots & \theta_p^0 \\ 0 & \theta_2^0 & \theta_3^0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \theta_p^0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}, \qquad W_2 = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ \theta_1^0 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \theta_{p-1}^0 & \theta_{p-2}^0 & \dots & 0 & 0 \\ \theta_p^0 & \theta_{p-1}^0 & \dots & \theta_1^0 & 0 \end{pmatrix},$$

$$S = \begin{pmatrix} 1 & -\theta_1^0 & \dots & -\theta_{p-1}^0 \\ 0 & 1 & \dots & -\theta_{p-2}^0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

Theorem 14.6 *If the bilinear model* (14.30) *is stationary*, $|W| \neq 0$, $|\varepsilon| \leq \varepsilon_+$, *and* $\varepsilon_+ \to 0$, *then the risk instability coefficient of the autoregressive forecasting statistic* (14.25) (*that uses the hypothetical model* (14.24) *with* $\varepsilon = 0$) *satisfies the asymptotic expansion*:

$$\kappa(\tau) = \varepsilon_+^2 \sigma^2 \left(1 + \bar{w}_{11} + \frac{(\sum_{j=1}^{\tau} \bar{s}_{1j})^2 + 4(1 - \sum_{i=1}^{p} \theta_i^0)^{-1} \sum_{j=1}^{\tau-1} \bar{s}_{1j} \bar{s}_{1,j+1}}{\sum_{j=1}^{\tau} \bar{s}_{1j}^2} \right) + o\big(\varepsilon_+^2\big).$$

For numerical illustration, consider the bilinear model $BL(10, 0, 1, 1)$ determined by (14.30) for $p = 10$: $x_t = -0.5x_{t-1} + 0.1x_{t-3} + 0.2x_{t-4} - 0.2x_{t-5} - 0.1x_{t-6} - 0.2x_{t-7} + 0.2x_{t-8} + 0.1x_{t-9} - 0.1x_{t-10} + \varepsilon x_{t-1} u_{t-1} + u_t$, $u_t \sim \mathcal{N}(0, 1)$. Figure 14.6 presents results of computer simulations on this model for two values of the forecasting horizon: $\tau = 1$ (one step ahead), $\tau = 5$ (five steps ahead); 10^3 Monte-Carlo simulation rounds were performed. Solid lines show the asymptotic values of the risk $r = \sigma^2 \sum_{j=1}^{\tau} \bar{s}_{1j}^2 (1 + \kappa^*(\tau))$ for $\tau = 1$ (lower curve) and for $\tau = 5$ (upper curve), where $\kappa^*(\tau)$ is the main term in the asymptotic expansion defined by Theorem 14.6; the circles near these curves indicate experimental values of the risk, and the dashed lines are 95 % confidence bounds.

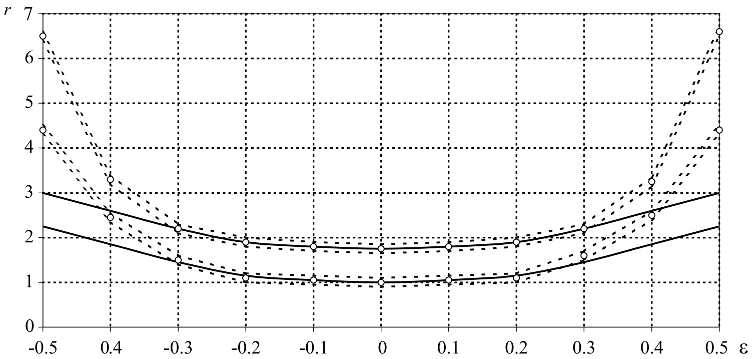


Fig. 14.6 Dependence of risk on the bilinearity coefficient

Table 14.1 The δ-admissible values $\varepsilon^* = \varepsilon^*(\delta, \tau)$

δ	τ		
	1	2	5
0.1	0.141	0.171	0.163
0.5	0.316	0.383	0.364
1.0	0.447	0.542	0.515

Table 14.1 presents δ-admissible values for the bilinearity coefficient $\varepsilon^* = \varepsilon^*(\delta, \tau)$.

14.6 Conclusions

The results presented in this paper provide a statistician with: (a) some classification scheme for typical distortions of the hypothetical models in statistical forecasting; (b) the guaranteed upper risk, the risk instability coefficient and the δ-admissible distortion level in "plug-in" statistical forecasting of time series under distorted regression and autoregression models; (c) the robust LM forecasting statistic under outliers and robust forecasting statistic under interval functional distortions of the hypothetical regression function. The theoretical results are tested on simulated and real statistical data; they are applied in the software package ROSTATFOR (RObust STATistical FORecasting) developed by the Belarusian State University.

Acknowledgements The Belarusian Statistical Association gratefully acknowledge Professor Ursula Gather for supporting of the International Conference "Computer Data Analysis and Modeling" in Minsk for many years by her invited lectures and work in the Program Committee. The author thanks anonymous Referees for the comments that have improved this Chapter.

The research was partially supported by the ISTC Project No. B-1910.

References

Anderson, T. W. (1971). *The statistical analysis of time series*. New York: Wiley.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. New York: Springer.

Bowerman, B., & O'Connel, R. T. (1993). *Forecasting and time series*. Belmont: Wadworth.

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis*. Oakland: Holden Day.

Davies, P. L., & Gather, U. (2004). Robust statistics. In *Handbook of computational statistics* (pp. 655–695). Berlin: Springer.

Davies, P. L., & Gather, U. (2005). Breakdown and groups (with discussion and rejoinder). *The Annals of Statistics*, *33*, 977–1035.

Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The Annals of Mathematical Statistics*, *34*, 447–456.

Fan, J., & Yao, Q. (2003). *Nonlinear time series*. New York: Springer.

Fried, R., & Gather, U. (2002). Fast and robust filtering of time series with trends. In W. Härdle & B. Rönz (Eds.), *Proceedings in computational statistics COMPSTAT* (pp. 367–372). Heidelberg: Physica.

Gather, U., & Kale, B. R. (1992). Outlier generating models—a review. In N. Venugopal (Ed.), *Contributions to stochastics* (pp. 57–85). New Delhi: Wiley.

Gather, U., Kuhnt, S., & Mühlenstädt, T. (2011). Industrial statistics. In M. Lovric (Ed.), *International encyclopedia of statistical science, part 9* (pp. 660–662). Berlin: Springer.

Gather, U., Schettlinger, K., & Fried, R. (2006). Online signal extraction by robust linear regression. *Computational Statistics*, *21*, 33–51.

Gelper, S., Fried, R., & Croux, C. (2010). Robust forecasting with exponential and Holt–Winters smoothing. *Journal of Forecasting*, *29*, 285–300.

Hampel, F. R., Ronchetti, E. M., & Rousseeuw, P. J. (1986). *Robust statistics*. New York: Wiley.

Huber, P. J. (1974). Mathematical problems in robust statistics. In R. D. James (Ed.), *Proceedings of the international congress of mathematicians* (pp. 821–824).

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Kharin, A., & Shlyk, P. (2009). Robust multivariate Bayesian forecasting under functional distortions in the chi-square metric. *Journal of Statistical Planning and Inference*, *139*, 3842–3846.

Kharin, Yu. (1996). *Robustness in statistical pattern recognition*. Dordrecht: Kluwer.

Kharin, Yu. (2003). Robustness analysis in forecasting of time series. In R. Dutter, P. Filzmoser, U. Gather, & P. J. Rousseeuw (Eds.), *Developments in robust statistics*, Heidelberg: Physica.

Kharin, Yu. (2008). *Optimality and robustness in statistical forecasting*. Minsk: BSU (in Russian).

Kharin, Yu. (2011a). Robustness of mean square risk in forecasting of regression time series. *Communications in Statistics. Theory and Methods*, *40*, 2893–2906.

Kharin, Yu. (2011b). Optimality and robustness in statistical forecasting. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 1034–1037). Berlin: Springer.

Kharin, Yu., & Voloshko, V. (2011). Robust estimation of AR coefficients under simultaneously influencing outliers and missing values. *Journal of Statistical Planning and Inference*, *141*, 3276–3288.

Maclaren, A. (1946). *Consolidated Webster encyclopedic dictionary*. New York: CBP.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: theory and methods*. Chichester: Wiley.

Morgenthaler, S. (2007). A survey of robust statistics. *Statistical Methods & Applications. Journal of the Italian Statistical Society*, *15*, 271–293.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, *1*, 327–332.

Chapter 15
Finding Outliers in Linear and Nonlinear Time Series

Pedro Galeano and Daniel Peña

15.1 Introduction

Outliers, or discordant observations, can have a strong effect on the model building process for a given time series. First, outliers introduce bias in the model parameter estimates, and then, distort the power of statistical tests based on biased estimates. Second, outliers may increase the confidence intervals for the model parameters. Third, as a consequence of the previous points, outliers strongly influence predictions. There are two main alternatives to analyze and treat outliers in time series. First, robust procedures can be applied to obtain parameter estimates not affected by the presence of outliers. These robust estimates are then used to identify outliers by using the residuals of the fit. Second, diagnostic methods are useful to detect the presence of outliers by analyzing the residuals of the model fit through iterative testing procedures. Once the outliers have been found, their effects are jointly estimated with the model parameters, obtaining, as a by-product, robust model parameter estimates. In this paper we focus on diagnostic methods and refer to Chap. 8 of Maronna et al. (2006) for a detailed review of robust procedures for ARMA models and Muler and Yohai (2008) and Muler et al. (2009) for two recent references.

For linear models, Fox (1972) introduced additive outliers (AO), which affect a single observation, and innovative outliers (IO), which affect a single innovation, and proposed the use of likelihood ratio test statistics for testing for outliers in autoregressive models. Tsay (1986) proposed an iterative procedure to identify outliers, to remove their effects, and to specify a tentative model for the underlying process. Chang et al. (1988) derived likelihood ratio criteria for testing the existence of outliers of both types and criteria for distinguishing between them and proposed

P. Galeano (✉) · D. Peña
Departamento de Estadística, Universidad Carlos III de Madrid, C/Madrid, 126, 28903 Getafe, Madrid, Spain
e-mail: pedro.galeano@uc3m.es

D. Peña
e-mail: daniel.pena@uc3m.es

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_15, © Springer-Verlag Berlin Heidelberg 2013

an iterative procedure for estimating the time series parameters in ARIMA models. Tsay (1988) extended the previous findings to two new types of outliers: the level shift (LS), which is a change in the level of the series, and the temporary change (TC), which is a change in the level of the series that decreases exponentially. Chen and Liu (1993a) proposed an iterative outlier detection to obtain joint estimates of model parameters and outlier effects that leads to more accurate model parameter estimates than previous ones. Luceño (1998) developed a multiple outlier detection method in time series generated by ARMA models based on reweighed maximum likelihood estimation. Gather et al. (2002) proposed a partially graphical procedure based on mapping the time series into a multivariate Euclidean space which can be applied online. Sánchez and Peña (2010) proposed a procedure that keeps the powerful features of previous methods but improves the initial parameter estimate, avoids confusion between innovative outliers and level shifts and includes joint tests for sequences of additive outliers in order to solve the masking problem. Finally, papers dealing with seasonal ARIMA models are Perron and Rodríguez (2003), Haldrup et al. (2011) and Galeano and Peña (2012), among others.

Recently, the focus has moved to outliers in nonlinear time series models. For instance, Chen (1997) proposed a method for detecting additive outliers in bilinear time series. Battaglia and Orfei (2005) proposed a model-based method for detecting the presence of outliers when the series is generated by a general nonlinear model that includes as particular cases the bilinear, the self-exciting threshold autoregressive (SETAR) model and the exponential autoregressive model, among others. In financial time series modeling, Doornik and Ooms (2005) presented a procedure for detecting multiple AO's in generalized autoregressive conditional heteroskedasticity (GARCH) models at unknown dates based on likelihood ratio test statistics. Carnero et al. (2007) studied the effect of outliers in the identification and estimation of GARCH models. Grané and Veiga (2010) proposed a general detection and correction method based on wavelets that can be applied to a large class of volatility models. Hotta and Tsay (2012) introduced two types of outlier in GARCH models: the level outlier (LO) corresponds to the situation in which a gross error affects a single observation that does not enter into the volatility equation, while the volatility outlier (VO) corresponds to the previous situation but the outlier enters into the volatility affecting all the remaining observations in the time series. Finally, Fokianos and Fried (2010) introduced three different outliers for the particular case of integer-valued GARCH (INGARCH) models and proposed a multiple outlier detection procedure for such outliers.

The literature on outliers in multivariate time series is brief. Tsay et al. (2000) generalized the four types of outliers usually considered in ARIMA models to the case of vector autoregressive moving average (VARMA) models and highlighted the differences between univariate and multivariate outliers. Importantly, the effect of a multivariate outlier not only depends on the model and the outlier size, as in the univariate case, but on the interaction between the model and size. These authors also proposed an iterative procedure for estimating the location, type and size of multivariate outliers. Galeano et al. (2006) proposed a method based on projections for identifying outliers without requiring initial specification of the multivariate model.

These authors showed that a multivariate outlier produces at least a univariate outlier in almost every projected series, and by detecting the univariate outliers, it is possible to identify the multivariate ones. Baragona and Battaglia (2007a) and Baragona and Battaglia (2007b) have proposed methods to discover outliers in dynamic factor models and in multiple time series by means of an independent component approach, respectively. Finally, Pankratz (1993) has considered outliers in dynamic regression models.

Other interesting issues related with outliers have been analyzed in the literature. For instance, detection of outliers in online monitoring data have been developed in Davies et al. (2004) and Gelper et al. (2009), among others. The effects of outliers in exponential smoothing techniques have been considered by Kirkendall (1992) and Koehler et al. (2012). The relationship between outliers, missing observations and interpolation techniques have been analyzed in Peña and Maravall (1991), Battaglia and Baragona (1992), Ljung (1993) and Baragona (1998). Forecasting time series with outliers have been addressed by Chen and Liu (1993b), for ARMA models, Franses and Ghijsels (1999), for GARCH models, and Gagné and Duchesne (2008), for dynamic vector time series models.

The rest of this contribution is organized as follows. In Sect. 15.2, we review outliers in univariate ARIMA models and discuss procedures for outlier detection and robust estimation. In Sect. 15.3, we consider outliers in non-linear time series models. Section 15.4 is devoted to outliers in multivariate time series models. Finally, Sect. 15.5 concludes the paper.

15.2 Outliers in ARIMA Models

This section reviews outliers in ARIMA time series models. We first introduce the four types of outliers usually considered in these models: additive outlier, innovative outlier, level shift and temporary change. Another type of unexpected events can be considered in the framework of an intervention event in the time series data, such as the ramp shift. Then, we describe procedures for outlier identification and estimation.

15.2.1 Types of Outliers in ARIMA Models

15.2.1.1 The ARIMA Model

We say that x_t follows an ARIMA(p,d,q) model if x_t can be written as:

$$\phi(B)(1-B)^d x_t = c + \theta(B)e_t, \tag{15.1}$$

where c is a constant, B is the backshift operator such that $Bx_t = x_{t-1}$, $\phi(B)$ and $\theta(B)$ are polynomials in B of orders p and q given by $\phi(B) = 1-\phi_1 B-\cdots-\phi_p B^p$

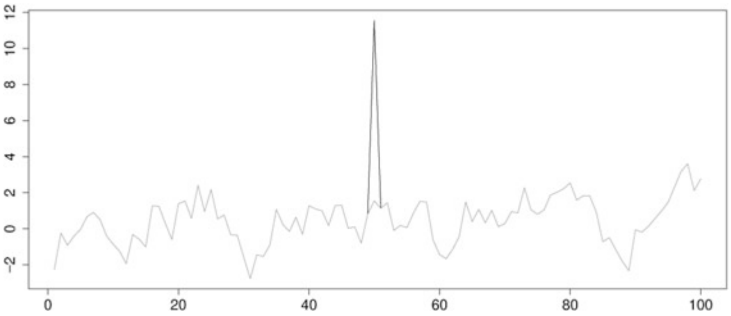


Fig. 15.1 Stationary series with and without an AO

and $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$, respectively, d is the number of unit roots and e_t is a white noise sequence of independent and identically distributed (i.i.d.) Gaussian with mean zero and variance σ_e^2. It is further assumed that the roots of $\phi(B)$ and $\theta(B)$ are outside the unit circle and have no common roots. The autoregressive representation of the ARIMA model in (15.1) is given by $\pi(B)x_t = c_\pi + e_t$, where $c_\pi = \theta^{-1}(B)c$ and $\pi(B) = \theta(B)^{-1}\phi(B)(1-B)^d$, while the moving average representation reduces to $x_t = c_\psi + \psi(B)e_t$, where $c_\psi = \phi^{-1}(B)(1-B)^{-d}c$ and $\psi(B) = \phi(B)^{-1}(1-B)^{-d}\theta(B)$.

15.2.1.2 Additive Outliers

An additive outlier (AO) corresponds to an exogenous change of a single observation of the time series and is usually associated with isolated incidents like measurement errors or impulse effects due to external causes. A time series $y_1, \ldots, y_T$ affected by the presence of an AO at $t = k$ is given by:

$$y_t = x_t + w I_t^{(k)}$$

for $t = 1, \ldots, T$, where x_t follows an ARIMA model in (15.1), w is the outlier size and $I_t^{(k)}$ is an indicator variable such that $I_t^{(k)} = 1$, if $t = k$, and $I_t^{(k)} = 0$, otherwise.

Figure 15.1 shows a simulated series with sample size $T = 100$ following an AR(1) model with parameter $\phi = 0.8$ and innovation variance $\sigma_e^2 = 1$, and the same series with an AO of size $w = 10$ at $t = 50$. Note how only a single observation is affected. An AO can have pernicious effects in all the steps of the time series analysis, i.e., model identification, estimation and prediction. For instance, the autocorrelation and partial autocorrelation functions, that are frequently used for model identification, can be severely affected by the presence of an AO.

15.2.1.3 Innovative Outliers

An innovative outlier (IO) corresponds to an endogenous change of a single innovation of the time series and is usually associated with isolated incidents like impulse

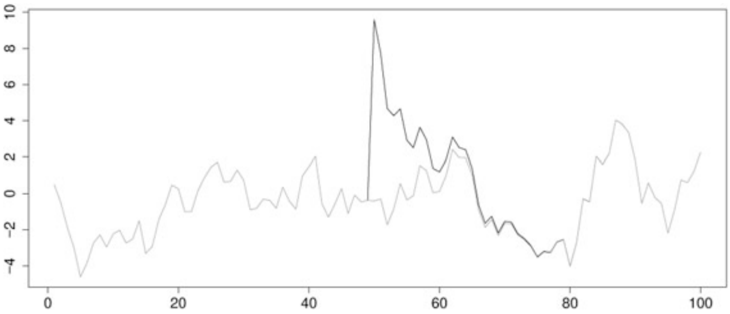


Fig. 15.2 Stationary series with and without an IO

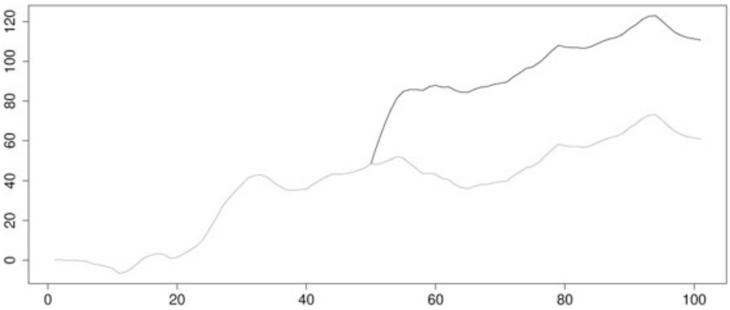


Fig. 15.3 Nonstationary series with and without an IO

effects due to internal causes. The innovations of a time series $y_1, \ldots, y_T$ affected by the presence of an IO at time point $t = k$ is given by:

$$a_t = e_t + wI_t^{(k)}, \tag{15.2}$$

where e_t are the innovations of the clean series x_t. Multiplying $\psi(B)$ in both sides of (15.2) leads to the equation for the observed series:

$$y_t = x_t + \psi(B)wI_t^{(k)}.$$

The effects of an IO on a series depend on the series being stationary or not. To see this point, Fig. 15.2 shows a simulated series with sample size $T = 100$ following an AR(1) model with parameter $\phi = 0.8$ and innovation variance $\sigma_e^2 = 1$, and the same series with an IO of size $w = 10$ at $t = 50$. Note how the IO modifies several observations of the series although its effect tends to disappear after a few observations. On the other hand, Fig. 15.3 shows a simulated series with sample size $T = 100$ following an ARIMA(1, 1, 0) model with parameter $\phi = 0.8$ and innovation variance $\sigma_e^2 = 1$, and the same series with an IO of size $w = 10$ at $t = 50$. Note how, in this case, the IO affects all the observations of the series starting from time point $t = 50$.

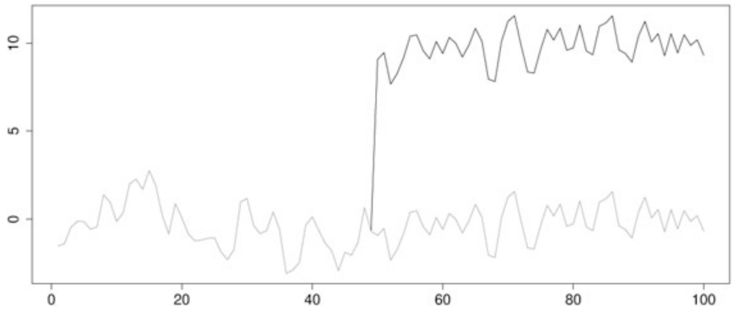


Fig. 15.4 Stationary series with and without a LS

15.2.1.4 Level Shifts

A level shift (LS) is a change in the mean level of the time series starting at $t = k$ and continuing until the end of the observed period. Therefore, a time series $y_1, \ldots, y_T$ affected by the presence of a LS at $t = k$ is given by:

$$y_t = x_t + wS_t^{(k)},$$

where $S_t^{(k)} = (1 - B)^{-1} I_t^{(k)}$ is a step function. Note that a LS serially affects the innovations as follows:

$$a_t = e_t + \pi(B)wS_t^{(k)}.$$

Figure 15.4 shows a simulated series with sample size $T = 100$ following an AR(1) model with parameter $\phi = 0.8$ and innovation variance $\sigma_e^2 = 1$, joint with the same series with a LS of size $w = 10$ at $t = 50$. Note how the LS affects all the observation of the series after $t = 50$. A LS has a strong effect in both identification and estimation of the observed series. Indeed, the effect of an LS is close to the effect of an IO on a nonstationary series.

15.2.1.5 Temporary Changes

A temporary change (TC) is a change with effect that decreases exponentially. Therefore, a time series $y_1, \ldots, y_T$ affected by the presence of a TC at $t = k$ is given by

$$y_t = x_t + \frac{1}{1 - \delta B} w I_t^{(k)},$$

where δ is the exponential decay parameter such that $0 < \delta < 1$. Note that if δ tends to 0, the TC reduces to an AO, whereas if δ tends to 1, the TC reduces to a LS. Under the presence of a TC, the innovations are affected as follows:

$$a_t = e_t + \frac{\pi(B)}{1 - \delta B} w I_t^{(k)}.$$

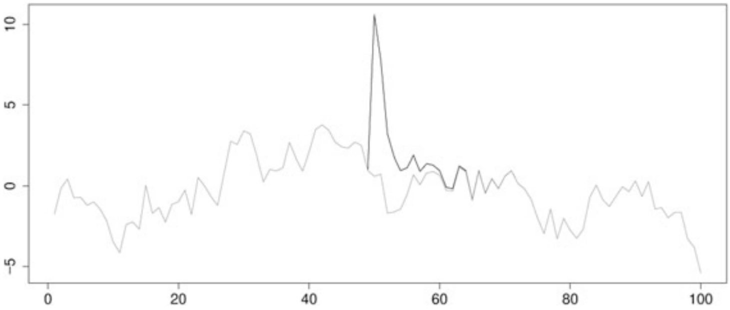


Fig. 15.5 Stationary series with and without a TC

Then, if $\pi(B)$ is close to $1-\delta B$, the effect of a TC on the innovations is very close to the effect of an IO. Otherwise, the TC can affect several innovations.

Figure 15.5 shows a simulated series with sample size $T=100$ following an AR(1) model with parameter $\phi=0.8$ and innovation variance $\sigma_e^2=1$, and the same series with an TC of size $w=10$ and decay rate $\delta=0.7$ at $t=50$. Note how the TC have a decreasing effect in the observations of the series after $t=50$.

15.2.1.6 Ramp Shifts

Finally, a ramp shift (RS) is a change in the trend of the time series in an ARIMA$(p,1,q)$ model starting at $t=k$ and continuing until the end of the observed period. Therefore, a time series $y_1,\ldots,y_T$ affected by the presence of a RS at $t=k$ is given by

$$y_t = x_t + wR_t^{(k)},$$

where $R_t^{(k)} = (1-B)^{-1}S_t^{(k)}$ is a ramp function. Note that a RS on the $I(1)$ series y_t is a LS on the differenced series $(1-B)y_t$. Note also that a RS serially affects the innovations as follows:

$$a_t = e_t + \pi(B)wR_t^{(k)}.$$

Other types of unexpected events have been considered in the literature. For instance, variance changes have been considered in Tsay (1988), while patches of additive outliers have been studied in Justel et al. (2001) and Penzer (2007). Modeling alternative unexpected events is also possible as the intervention framework is flexible enough to model many different situations. For example, new effects can be defined using combinations of the outliers previously considered.

15.2.2 Outlier Identification and Estimation

In general, all the types of outliers that we have presented can be written in a general equation:

$$y_t = x_t + \nu(B) w I_t^{(k)}, \tag{15.3}$$

where $\nu(B) = 1$ for an AO, $\nu(B) = \psi(B)$ for an IO, $\nu(B) = (1 - B)^{-1}$ for a LS, $\nu(B) = (1 - \delta B)^{-1}$ for a TC and $\nu(B) = (1 - B)^{-2}$ for a RS, respectively. Therefore, outliers in time series can be seen as particular cases of interventions, introduced by Box and Tiao (1975), to model dynamic changes on a time series at known time points.

Assume that we observe a series, $y_1, \dots, y_T$, following an ARIMA(p, d, q) model as in (15.1) with known parameters and with an outlier of known type at $t = k$. Multiplying by $\pi(B)$ in (15.3) leads to the equation for the innovations:

$$a_t = e_t + w_i z_{i,t}, \tag{15.4}$$

for $i =$ AO, IO, LS and TC, where w_{AO}, w_{IO}, w_{LS} and w_{TC} is the size of the outlier for AO, IO, LS and TC, respectively, and $z_{i,t} = \pi(B)\nu_i(B) I_t^{(k)}$, where $\nu_{\mathrm{AO}}(B) = 1$, $\nu_{\mathrm{IO}}(B) = \psi(B)$, $\nu_{\mathrm{LS}}(B) = (1 - B)^{-1}$ and $\nu_{\mathrm{TC}}(B) = (1 - \delta B)^{-1}$, respectively. From (15.4), for any particular case, one can easily estimate the size of the outlier by least-squares leading to:

$$\hat{w}_i = \frac{\sum_{t=1}^T z_{i,t} a_t}{\sum_{t=1}^T z_{i,t}^2}$$

with variance $\rho_i^2 \sigma_e^2$ where $\rho_i^2 = (\sum_{t=1}^T z_{i,t}^2)^{-1}$. Consequently, knowing the type and location of the outlier, it is easy to adjust the outlier effect on the observed series using the corresponding estimates, $\hat{w}_{\mathrm{AO}}$, $\hat{w}_{\mathrm{IO}}$, $\hat{w}_{\mathrm{LS}}$ or $\hat{w}_{\mathrm{TC}}$, respectively.

Also, the estimates of the outlier size can now be used to test whether one outlier of known type has occurred at $t = k$. Indeed, the likelihood ratio test statistic for the null hypothesis $H_0 : w_i = 0$ against the alternative $H_1 : w_i \neq 0$, is given by

$$\tau_{i,k} = \frac{\hat{w}_{i,k}}{\rho_i \sigma_e}.$$

The statistic $\tau_{i,k}$, under the null hypothesis, follows a Gaussian distribution.

However, in practice, the number, location, type and size of the outliers are unknown. Several papers, including Chang et al. (1988), Tsay (1988), Chen and Liu (1993a) and Sánchez and Peña (2010), among others, have proposed iterative procedures in which the idea is to compute the likelihood ratio test statistics for all the observations of the series under the null hypothesis of no outliers. In particular, the procedure by Chen and Liu (1993a), which is standard nowadays, works as follows. In a first step, an ARIMA model is identified for the series and the parameters are estimated using maximum likelihood. Then, the likelihood ratio test statistics $\tau_{i,t}$, for

$i =$ AO, IO, LS and TC, are computed. If the maximum of all these statistics is significant, an outlier of the type that provides with the maximum statistic is detected. Then, the series is cleaned of the outlier effects and the parameters of the model are re-estimated. This step is repeated until no more outliers are found. In a second step, the outliers effects and the ARIMA model parameters are estimated jointly using a multiple regression model. If some outlier is not significant, it is removed from the outliers set. Then, the multiple regression model is re-estimated. This step is repeated until all the outlier effects are significant. In a final step, the two previous steps are repeated but initially using the ARIMA model parameters estimates obtained at the end of the second step. However, this procedure has three main drawbacks. First, when a level shift is present in the series, the procedure tends to identify an innovative outlier instead of the level shift. Second, the initial estimation of the model parameters usually leads to a very biased set of parameters that may produce the procedure to fail. Third, the masking and swamping effects, although mitigated with respect to previous procedures, are still present if a sequence of outlier patches is present in the time series. Sánchez and Peña (2010) proposed a procedure for multiple outlier detection and robust estimation that tries to avoid these three problems. In particular, to solve the first problem, it is proposed to compare AO versus IO and deal with LS alone. To solve the second problem, it is proposed to use influence measures to identify the observations that have a larger impact on estimation and estimate the parameters assuming that the most influential observations are missing. Finally, to solve the third problem, an influence measure for LS or sequences of patchy outliers is proposed that can be used to carry out the initial cleaning of the time series.

15.3 Outliers in Nonlinear Time Series Models

This section reviews outliers in some nonlinear time series models. In particular, we first consider the model-based method proposed by Battaglia and Orfei (2005) for detecting the presence of outliers when the series is generated by a general nonlinear model. Second, we summarize the effect of outliers in GARCH models following Carnero et al. (2007) and present a method proposed by Hotta and Tsay (2012) for detecting outliers in GARCH models. Finally, we describe the method proposed by Fokianos and Fried (2010) to detect outliers in INGARCH models.

15.3.1 Outliers in a General Nonlinear Model

Battaglia and Orfei (2005) assumed a time series x_t following the model:

$$x_t = f\left(\mathbf{x}^{(t-1)}, \mathbf{e}^{(t-1)}\right) + e_t, \tag{15.5}$$

where f is a nonlinear function also containing unknown parameters, $\mathbf{x}^{(t-1)} = (x_{t-1}, x_{t-2}, \ldots, x_{t-p})'$, $\mathbf{e}^{(t-1)} = (e_{t-1}, e_{t-2}, \ldots, e_{t-p})'$ and e_t is a white noise sequence of independent and identically distributed (i.i.d.) Gaussian with mean zero and variance σ_e^2. Note that the model in (15.5) covers several well known nonlinear models, such as the bilinear model, the self-exciting threshold autoregressive (SETAR) model and the exponential autoregressive model, among others.

For the model in (15.5), Battaglia and Orfei (2005) consider additive and innovative outliers. First, for an AO at $t = k$, the observed series is $y_1, \ldots, y_T$, given by $y_t = x_t + wI_t^{(k)}$, for $t = 1, \ldots, T$, where x_t follows the model in (15.5). Therefore, the observed series can be written as $y_t = f(\mathbf{y}^{(t-1)}, \mathbf{a}^{(t-1)}) + a_t$, where $\mathbf{y}^{(t-1)} = (y_{t-1}, y_{t-2}, \ldots, y_{t-p})'$ and $\mathbf{a}^{(t-1)} = (a_{t-1}, a_{t-2}, \ldots, a_{t-p})'$, respectively, for $t = 1, \ldots, T$. The innovations of the observed series can be obtained recursively from $a_t = y_t - f(\mathbf{y}^{(t-1)}, \mathbf{a}^{(t-1)})$. On the other hand, for an IO at $t = k$, the observed series is given by $y_t = f(\mathbf{y}^{(t-1)}, \mathbf{a}^{(t-1)}) + a_t$, where $a_t = e_t + wI_t^{(k)}$, where $\mathbf{y}^{(t-1)} = (y_{t-1}, y_{t-2}, \ldots, y_{t-p})'$ and $\mathbf{a}^{(t-1)} = (a_{t-1}, a_{t-2}, \ldots, a_{t-p})'$, respectively, for $t = 1, \ldots, T$.

Estimation of outlier effects can be done similarly to the ARIMA case through least squares. Battaglia and Orfei (2005) showed that the LS estimate of w for an IO is given by $\hat{w}_{\mathrm{IO}} = a_k$ with variance σ_e^2, and that the LS estimate of w for an AO is given by

$$\hat{w}_{\mathrm{AO}} = \frac{\sum_{j=0}^{T-k} c_j a_{k+j}}{\sum_{j=0}^{T-k} c_j^2}$$

with variance $(\sum_{j=0}^{T-k} c_j^2)\sigma_e^2$, where

$$c_j = -\left[\frac{\partial}{\partial y_{t-j}} f\left(y^{(k+j-1)}, a^{(k+j-1)}\right) + \sum_{i=1}^{j} c_{j-i} \frac{\partial}{\partial a_{t-j}} f\left(y^{(k+j-1)}, a^{(k+j-1)}\right) \right],$$

for $j = 1, \ldots, T - k$. Consequently, the likelihood ratio test statistics to test for the presence of an AO and a IO at $t = k$ are given by

$$\tau_{\mathrm{IO}} = \frac{a_k}{\sigma_e},$$

and

$$\tau_{\mathrm{AO}} = \frac{\sum_{j=0}^{T-k} c_j a_{k+j}}{\sqrt{\sum_{j=0}^{T-k} c_j^2}\sigma_e},$$

respectively. Under the null hypothesis of no outlier, τ_{AO} and τ_{IO} have a standard Gaussian distribution.

In order to detect the presence of several outliers in a nonlinear time series, Battaglia and Orfei (2005) considered a procedure similar to that used in Chen and Liu (1993a).

15.3.2 Outliers in GARCH Models

Carnero et al. (2007) have analyzed the effects of outliers on the identification and estimation of GARCH models. Regarding identification, Carnero et al. (2007) derived the asymptotic biases caused by outliers on the sample autocorrelations of squared observations generated by stationary processes and obtained the asymptotic biases of the ordinary least squares (OLS) estimator of the parameters of ARCH(p) models. Finally, these authors also studied the effects of outliers on the estimated asymptotic standard deviations of the estimators considered and showed that they are biased estimates of the sample standard deviations.

Recently, Hotta and Tsay (2012) have distinguished two types of outliers in GARCH models and have proposed a method for their detection. For simplicity of presentation, we consider the ARCH(1) model given by

$$\begin{aligned} x_t &= \sqrt{h_t} e_t, \\ h_t &= \alpha_0 + \alpha_1 x_{t-1}^2, \end{aligned}$$

where $\alpha_0 > 0$, $0 < \alpha_1 < 1$, and e_t are independent and identically distributed standard Gaussian random variables. Outliers in an ARCH(1) model encounter two different scenarios because an outlier can affect the level of x_t or the volatility h_t. Therefore, a volatility outlier, denoted by VO, and defined as follows

$$\begin{aligned} y_t &= \sqrt{h_t} e_t + w I_t^{(k)}, \\ h_t &= \alpha_0 + \alpha_1 y_{t-1}^2 \end{aligned}$$

affects the volatility of the series, while a level outlier, denoted by LO, and given by

$$\begin{aligned} y_t &= \sqrt{h_t} e_t + w I_t^{(k)}, \\ h_t &= \alpha_0 + \alpha_1 \big(y_{t-1} - w I_{t-1}^{(k)}\big)^2 \end{aligned}$$

only affects the level of the series at the observation where it occurs.

Hotta and Tsay (2012) estimated w by means of the ML estimation method. These authors showed that the ML estimate of w for a VO is given by

$$\hat{w}_{\mathrm{VO}} = y_k,$$

and that there are two ML estimates of w for a LO: the first one is $\hat{w}_{\mathrm{LO}} = y_k$ and the second one is $\hat{w}_{\mathrm{LO}} = y_k - \hat{x}_k$, where $\hat{x}_k =$ is the square root of the positive solution of the second order equation:

$$\begin{aligned} g(x) = {} & \alpha_1^2 x^2 + \big(2\alpha_0\alpha_1 + \alpha_1^2\big(\alpha_0 + \alpha_1 y_{k-1}^2\big)\big)x \\ & + \big(\alpha_0^2 + \alpha_0\alpha_1\big(\alpha_0 + \alpha_1 y_{k-1}^2\big) - \alpha_1 y_{k+1}^2\big(\alpha_0 + \alpha_1 y_{k-1}^2\big)\big), \end{aligned}$$

provided that such a solution exists.

In order to test for the presence of a VO or a LO, Hotta and Tsay (2012) propose the use of the Lagrange multiplier (LM) test statistic that, for a VO, is defined as

$$\mathrm{LM}_k^{\mathrm{VO}} = \frac{y_k^2}{\alpha_0 + \alpha_1 y_{k-1}^2},$$

while for a LO, it is given by

$$\mathrm{LM}_k^{\mathrm{LO}} = \mathrm{LM}_k^{\mathrm{VO}} \left\{ 1 + \alpha_1 h_k \left(\frac{1}{h_{k+1}} - \frac{y_{k+1}^2}{h_{k+1}^2} \right) \right\}^2 \left(1 + 2\alpha_1^2 h_k \frac{y_k^2}{h_{k+1}^2} \right)^{-1}.$$

Note that $\mathrm{LM}_k^{\mathrm{LO}} = \mathrm{LM}_k^{\mathrm{VO}}$ if $\alpha_1 = 0$. Therefore, the two test statistics should not differ substantially when α_1 is close to 0. To detect multiple outliers, Hotta and Tsay (2012) thus propose to compute the maximum LM statistics

$$\mathrm{LM}_{\max}^{\mathrm{VO}} = \max_{2 \le t \le n} \mathrm{LM}_k^{\mathrm{VO}}, \qquad \mathrm{LM}_{\max}^{\mathrm{LO}} = \max_{2 \le t \le n} \mathrm{LM}_k^{\mathrm{LO}},$$

for which it is easy to compute critical values via simulation. If both statistics are significant, one may choose the outlier that gives the smaller p-value.

15.3.3 Outliers in INGARCH Models

Fokianos and Fried (2010) consider outliers in the integer-valued GARCH (INGARCH) model given by

$$\begin{aligned} x_t \mid \mathcal{F}_{t-1}^x &\sim \text{Poisson}(\lambda_t), \\ \lambda_t &= \alpha_0 + \sum_{j=1}^{p} \alpha_j \lambda_{t-j} + \sum_{i=1}^{q} \beta_i x_{t-i}, \end{aligned} \tag{15.6}$$

for $t \ge 1$, where λ_t is the Poisson intensity of the process x_t, $\mathcal{F}_{t-1}^x$ stands for the σ-algebra generated by $\{x_{t-1}, \ldots, x_{1-q}, \lambda_{t-1}, \ldots, \lambda_0\}$, α_0 is an intercept, $\alpha_j > 0$, for $j = 1, \ldots, p$ and $\beta_i > 0$, for $i = 1, \ldots, q$ and $\sum_{j=1}^{p} \alpha_j + \sum_{i=1}^{q} \beta_i < 1$ to get covariance stationarity. Outliers in the INGARCH model (15.6) can be written as

$$\begin{aligned} y_t \mid \mathcal{F}_{t-1}^y &\sim \text{Poisson}(\kappa_t), \\ \kappa_t &= \alpha_0 + \sum_{j=1}^{p} \alpha_j \kappa_{t-j} + \sum_{i=1}^{q} \beta_i y_{t-i} + w(1 - \delta B)^{-1} I_t^{(k)}, \end{aligned} \tag{15.7}$$

for $t \ge 1$, where κ_t is the Poisson intensity of the process y_t, $\mathcal{F}_{t-1}^y$ stands for the σ-algebra generated by $\{y_{t-1}, \ldots, y_{1-q}, \kappa_{t-1}, \ldots, \kappa_0\}$, w is the size of the outlier, and $0 \leq \delta \leq 1$ is a parameter that controls the outlier effect. In particular, $\delta = 0$ corresponds to an spiky outlier (SO) that influences the process from time k on, but to a rapidly decaying extent provided that α_1 is not close to unity, $0 < \delta < 1$ corresponds to a transient shift (TS) that affects several consecutive observations although its effect becomes gradually smaller as time grows and, finally, $\delta = 1$ corresponds to a level shift (LS) that affects permanently the mean and the variance of the observed series.

Fokianos and Fried (2010) propose to estimate the outlier effect w via conditional maximum likelihood. Therefore, given the observed time series $y_1, \ldots, y_T$, the log-likelihood of the parameters of model (15.7), $\boldsymbol{\eta} = (\alpha_0, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q, w)'$ conditional on $\mathcal{F}_0^y$ is given, up to a constant, by

$$\ell(\boldsymbol{\eta}) = \prod_{t=q+1}^{T} \big(y_t \log \kappa_t(\boldsymbol{\eta}) - \kappa_t(\boldsymbol{\eta})\big) \tag{15.8}$$

with score function

$$\frac{\partial l(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \sum_{t=q+1}^{T} \left(\frac{y_t}{\kappa_t(\boldsymbol{\eta})} - 1\right) \frac{\partial \kappa_t(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}.$$

In addition, the conditional information matrix for $\boldsymbol{\eta}$ is given by

$$\mathbf{G}(\boldsymbol{\eta}) = \sum_{t=q+1}^{T} \mathrm{Cov}\left[\frac{\partial l(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \,\middle|\, \mathcal{F}_{t-1}^y\right] = \sum_{t=1}^{T} \frac{1}{\kappa_t(\boldsymbol{\eta})} \left(\frac{\partial l(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right)\left(\frac{\partial l(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right)'.$$

Consequently, $\hat{w}$ is obtained from the ML estimate $\hat{\boldsymbol{\eta}}$ after maximizing the log-likelihood function (15.8). In order to test for the presence of an outlier at $t = k$, Fokianos and Fried (2010) propose to use the score test given by

$$T_k = \Delta' \mathbf{G}(\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_p, \tilde{\beta}_1, \ldots, \tilde{\beta}_q, 0)^{-1} \Delta,$$

where

$$\Delta = \frac{\partial l(\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_p, \tilde{\beta}_1, \ldots, \tilde{\beta}_q, 0)}{\partial \boldsymbol{\eta}},$$

and $(\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_p, \tilde{\beta}_1, \ldots, \tilde{\beta}_q, 0)'$ is the vector that contains the ML estimates of the parameters of the model (15.6) and the value $w = 0$. Under the null hypothesis of no outlier, T_k has an asymptotic χ_1^2 distribution. To detect an outlier of a certain type at an unknown time point, the idea is to obtain

$$T = \max_{q+1 \leq t \leq T} T_t,$$

and reject the null hypothesis of no outlier if T is large. The distribution of this statistic can be calibrated using bootstrap. Finally, to detect the presence of several outliers in a INGARCH time series, Fokianos and Fried (2010) proposed a procedure similar to that used in Chen and Liu (1993a).

15.4 Outliers in Multivariate Time Series Models

Outliers in multivariate time series has been much less analyzed than in the univariate case. Multivariate outliers were introduced in Tsay et al. (2000). These authors have also proposed a detection method based on individual and joint likelihood ratio statistics. Alternatively, Galeano et al. (2006) used projection pursuit methods to develop a procedure for detecting outliers. In particular, Galeano et al. (2006) showed that testing for outliers in certain projection directions can be more powerful than testing the multivariate series directly. In view of these findings, an iterative procedure to detect and handle multiple outliers based on a univariate search in these optimal directions were proposed. The main advantage of this procedure is that it can identify outliers without prespecifying a vector ARMA model for the data. An alternative method based on linear combinations of the components of the vector of time series can be found in Baragona and Battaglia (2007b) that considers an independent component approach. Finally, Baragona and Battaglia (2007a) have proposed a method to discover outliers in a dynamic factor model based on linear transforms of the observed time series. In this section, we briefly review the main findings in Tsay et al. (2000) and Galeano et al. (2006).

15.4.1 The Tsay, Peña and Pankratz Procedure

A r-dimensional vector time series $\mathbf{X}_t = (X_{1t}, \ldots, X_{rt})'$ follows a vector ARMA (VARMA) model if

$$\Phi(B)\mathbf{X}_t = \mathbf{C} + \Theta(B)\mathbf{E}_t, \quad t = 1, \ldots, T, \tag{15.9}$$

where $\Phi(B) = \mathbf{I} - \Phi_1 B - \cdots - \Phi_p B^p$ and $\Theta(B) = \mathbf{I} - \Theta_1 B - \cdots - \Theta_q B^q$ are $r \times r$ matrix polynomials of finite degrees p and q, $\mathbf{C}$ is a r-dimensional constant vector, and $\mathbf{E}_t = (E_{1t}, \ldots, E_{rt})'$ is a sequence of independent and identically distributed Gaussian random vectors with zero mean and positive-definite covariance matrix Σ. The autoregressive representation of the VARMA model in (15.9) is given by $\Pi(B)\mathbf{X}_t = \mathbf{C}_\Pi + \mathbf{E}_t$, where $\Phi(1)\mathbf{C}_\Pi = \mathbf{C}$ and $\Pi(B) = \Theta(B)^{-1}\Phi(B) = \mathbf{I} - \sum_{i=1}^{\infty} \Pi_i B^i$, while the moving-average representation of $\mathbf{X}_t$ is given by $\mathbf{X}_t = \mathbf{C}_\Psi + \Psi(B)\mathbf{E}_t$, where $\Phi(1)\mathbf{C}_\Psi = \mathbf{C}$ and $\Phi(B)\Psi(B) = \Theta(B)$ with $\Psi(B) = \mathbf{I} + \sum_{i=1}^{\infty} \Psi_i B^i$.

Tsay et al. (2000) generalize four types of univariate outliers to the vector case. Under the presence of a multivariate outlier, we observe a time series $\mathbf{Y} = (Y_1', \ldots, Y_T')'$, where $\mathbf{Y}_t = (Y_{1t}, \ldots, Y_{rt})'$, can be written as follows:

$$\mathbf{Y}_t = \mathbf{X}_t + \Lambda(B)\mathbf{w}I_t^{(k)}, \tag{15.10}$$

where $\mathbf{w} = (w_1, \ldots, w_r)'$ is the size of the outlier and $\mathbf{X}_t$ follows a VARMA model. The type of the outlier is defined by the matrix polynomial $\Lambda(B)$: for a multivariate innovational outlier (MIO), $\Lambda(B) = \Psi(B)$; for a multivariate additive outlier (MAO), $\Lambda(B) = \mathbf{I}$; for a multivariate level shift (MLS), $\Lambda(B) = (1 - B)^{-1}\mathbf{I}$;

and, finally, for a multivariate temporary (or transitory) change (MTC), $\Lambda(B) = (\mathbf{I} - \delta \mathbf{I} B)^{-1}$. In practice, an outlier may produce a complex effect, given by a linear combination of the previously discussed pure effects. Furthermore, different components of $\mathbf{X}_t$ may suffer different outlier effects. An example of this kind of mixed effects can be found in Galeano et al. (2006).

Given the parameters of the VARMA model for $\mathbf{X}_t$, the series of innovations are defined by $\mathbf{A}_t = \Pi(B)\mathbf{Y}_t - \mathbf{C}_\Pi$ and the relationship with the true innovations, $\mathbf{E}_t$, is given by

$$\mathbf{A}_t = \mathbf{E}_t + \Gamma(B)\mathbf{w} I_t^{(k)},$$

where $\Gamma(B) = \Pi(B)\Lambda(B) = \mathbf{I} - \sum_{i=1}^{\infty} \Gamma_i B^i$. Now, the least squares estimate of the size of an outlier of type i at time point k is given by

$$\mathbf{w}_{i,k} = -\left(\sum_{j=0}^{n-k} \Gamma_j' \Sigma^{-1} \Gamma_j \right)^{-1} \left(\sum_{j=0}^{n-k} \Gamma_j' \Sigma^{-1} \mathbf{A}_{k+j} \right),$$

where $\Gamma_0 = -\mathbf{I}$ and $i =$ MIO, MAO, MLS and MTC for subscripts, and has a covariance matrix given by $\Sigma_{i,k} = (\sum_{j=0}^{n-k} \Gamma_j' \Sigma^{-1} \Gamma_j)^{-1}$. The likelihood ratio test statistic for testing for the presence of a multivariate outlier of type i at $t = k$ is $J_{i,k} = \mathbf{w}_{i,k}' \Sigma_{i,k}^{-1} \mathbf{w}_{i,k}$. Under the null hypothesis of no outlier, $J_{i,k}$ has a χ_r^2 distribution. Tsay et al. (2000) also proposed a second statistic defined by $C_{i,k} = \max\{|w_{j,i,k}|/\sqrt{\sigma_{j,i,k}} : 1 \le j \le r\}$, where $w_{j,i,k}$ is the jth element of $\mathbf{w}_{i,k}$ and $\sigma_{j,i,k}$ is the jth element of the main diagonal of $\Sigma_{i,k}$, with the aim of look for outliers in individual components of the vector of series.

In practice, the parameter matrices are then substituted by their estimates and the following overall test statistics are defined:

$$J_{\max}(i, k_i) = \max_{1 \le t \le n} J_{i,t}, \qquad C_{\max}\left(i, k_i^*\right) = \max_{1 \le t \le n} C_{i,t},$$

where k_i and k_i^* denote respectively the time points at which the maximum of the joint test statistics and the maximum component statistics occur.

15.4.2 The Galeano, Peña and Tsay Procedure

Galeano et al. (2006) have proposed a method for detecting multivariate outliers in time series without requiring initial specification of the multivariate model. This is very important in these settings because model identification is quite complicated in the presence of outliers. The method is based on univariate outlier detection applied to some useful projections of the vector time series. The basic idea is simple: a multivariate outlier produces at least a univariate outlier in almost every projected series, and by detecting the univariate outliers we can identify the multivariate ones.

First, a non-zero linear combination of the components of the VARMA model in (15.9) follows a univariate ARMA model. Second, when the observed series $\mathbf{Y}_t$ is affected by an outlier, as in (15.10), the projected series $y_t = \mathbf{v}'\mathbf{Y}_t$ satisfies $y_t = x_t + \mathbf{v}'\Lambda(B)\mathbf{w}I_t^{(k)}$. Specifically, if $\mathbf{Y}_t$ has a MAO, the projected series is $y_t = x_t + \omega I_t^{(k)}$, so that it has an additive outlier of size $\omega = \mathbf{v}'\mathbf{w}$ at point $t = k$ provided that $\mathbf{v}'\mathbf{w} \neq 0$. Similarly, the projected series of a vector process with a MLS of size $\mathbf{w}$ will have a level shift with size $\omega = \mathbf{v}'\mathbf{w}$ at $t = k$. The same result also applies to MTC. A MIO can produce several effects. In particular, a MIO can lead to a patch of consecutive outliers with sizes $\mathbf{v}'\mathbf{w}, \mathbf{v}'\Psi_1\mathbf{w}, \ldots, \mathbf{v}'\Psi_{T-h}\mathbf{w}$, starting at $t = k$. Assuming that k is not close to T and because $\Psi_j \to 0$, the size of the outlier in the patch tends to zero. In the particular case that $\mathbf{v}'\Psi_i\mathbf{w} = \psi_i\mathbf{v}'\mathbf{w}$, for $i = 1, \ldots, T-k$, then y_t has an innovational outlier at $t = k$ with size $\beta = \mathbf{v}'\mathbf{w}$. However, if $\mathbf{v}'\Psi_i\mathbf{w} = 0$, for $i = 1, \ldots, T-k$, then y_t has an additive outlier at $t = k$ with size $\mathbf{v}'\mathbf{w}$, and if $\mathbf{v}'\Psi_i\mathbf{w} = \mathbf{v}'\mathbf{w}$, for $i = 0, \ldots, T-k$, then y_t has a level shift at $t = k$ with size $\beta = \mathbf{v}'\mathbf{w}$. Therefore, the univariate series y_t obtained by the projection can be affected by an additive outlier, a patch of outliers or a level shift.

Galeano et al. (2006) have shown that it is possible to identify multivariate outliers better by applying univariate test statistics to optimal projections than by using multivariate statistics on the original series. More precisely, it is possible to show that, in the presence of a multivariate outlier, the directions that maximize or minimize the kurtosis coefficient of the projected series include the direction of the outlier, that is, the direction that maximizes the ratio between the outlier size and the variance of the projected observations. Therefore, Galeano et al. (2006) proposed here a sequential procedure for outlier detection based on the directions that minimize and maximize the kurtosis coefficient of the projections. The procedure is divided into four steps: (1) obtain the optimal directions; (2) search for outliers in the projected univariate time series; (3) remove the effect of all detected outliers by using an approximated multivariate model; (4) iterate the previous steps applied to the cleaned series until no more outliers are found. It is important to note that in Step (2), the detection is carried out in two stages: first, MLS's are identified; second, MIO's, MAO's and MTC's are found. This is done in order to avoid confusions between multivariate innovational outliers and multivariate level shifts.

15.5 Conclusions

This chapter summarized outliers in both univariate and multivariate time series. Although many work has been done, more research is still needed in order to analyze outliers and unexpected events in time series. First, new effects in nonlinear time series models can be considered. For instance, level shifts in bilinear and SETAR models, transitory changes in GARCH models or additive outliers in INGARCH models would be of interest. Second, as far as we know, outlier detection in multivariate nonlinear time series models have been not considered yet. For instance, the extension of additive and volatility outliers to most of the available multivariate

GARCH models is almost straightforward. Likewise, outliers in multivariate bilinear and SETAR time series models is of interest. Finally, most of the existing literature on outlier detection focus on iterative testing procedures. Recently, Galeano and Peña (2012) have proposed a method to detect additive outliers by means of the use of a model selection criterion. The main advantage of the procedure is that all the outliers are detected in a single step. Although the computational cost of the procedure is high, the detection of outliers by means of model selection criteria is a promising line of research. Finally, this chapter is closely related with those by Barme–Delcroix (Chap. 3) who analyzes extreme events and Huskova (Chap. 11) who analyzes robust change point analysis.

Acknowledgements We gratefully acknowledge financial support by MCI grants SEJ2007/64500 and MTM2008-03010.

References

Baragona, R. (1998). Nonstationary time series, linear interpolators and outliers. *Statistica*, *58*, 375–394.

Baragona, R., & Battaglia, F. (2007a). Outliers in dynamic factor models. *Electronic Journal of Statistics*, *1*, 392–432.

Baragona, R., & Battaglia, F. (2007b). Outliers detection in multivariate time series by independent component analysis. *Neural Computation*, *19*, 1962–1984.

Battaglia, F., & Orfei, L. (2005). Outlier detection and estimation in nonlinear time series. *Journal of Time Series Analysis*, *26*, 107–121.

Battaglia, F., & Baragona, R. (1992). Linear interpolators and the outliers problem in time series. *Metron*, *50*, 79–97.

Box, G. E. P., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, *70*, 70–79.

Carnero, M. A., Peña, D., & Ruiz, E. (2007). Effects of outliers on the identification and estimation of GARCH models. *Journal of Time Series Analysis*, *28*, 471–497.

Chang, I., Tiao, G. C., & Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, *30*, 193–204.

Chen, C. W. S. (1997). Detection of additive outliers in bilinear time series. *Computational Statistics & Data Analysis*, *24*, 283–294.

Chen, C., & Liu, L. M. (1993a). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, *88*, 284–297.

Chen, C., & Liu, L. M. (1993b). Forecasting time series with outliers. *Journal of Forecasting*, *12*, 13–35.

Davies, P. L., Fried, R., & Gather, U. (2004). Robust signal extraction for on-line monitoring data. *Journal of Statistical Planning and Inference*, *122*, 65–78.

Doornik, J. A., & Ooms, M. (2005). *Outlier detection in GARCH models*. 2005-W24 Nuffield economics working papers.

Fokianos, C., & Fried, R. (2010). Interventions in INGARCH processes. *Journal of Time Series Analysis*, *31*, 210–225.

Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society. Series B. Methodological*, *34*, 350–363.

Franses, P. H., & Ghijsels, H. (1999). Additive outliers, GARCH and forecasting volatility. *International Journal of Forecasting*, *15*, 1–9.

Gagné, C., & Duchesne, P. (2008). On robust forecasting in dynamic vector time series models. *Journal of Statistical Planning and Inference*, *138*, 3927–3938.

Galeano, P., & Peña, D. (2012). Additive outlier detection in seasonal ARIMA models by a modified Bayesian information criterion. In W. R. Bell, S. H. Holan, & T. S. McElroy (Eds.), *Economic time series: modeling and seasonality* (pp. 317–336). Boca Raton: Chapman & Hall.

Galeano, P., Peña, D., & Tsay, R. S. (2006). Outlier detection in multivariate time series by projection pursuit. *Journal of the American Statistical Association*, *101*, 654–669.

Gather, U., Bauer, M., & Fried, R. (2002). The identification of multiple outliers in online monitoring data. *Estatistica*, *54*, 289–338.

Gelper, S., Schettlinger, K., Croux, C., & Gather, U. (2009). Robust online scale estimation in time series: a model-free approach. *Journal of Statistical Planning and Inference*, *139*, 335–349.

Grané, A., & Veiga, H. (2010). Wavelet-based detection of outliers in financial time series. *Computational Statistics & Data Analysis*, *54*, 2580–2593.

Haldrup, N., Montañes, A., & Sansó, A. (2011). Detection of additive outliers in seasonal time series. *Journal of Time Series Econometrics*, *3*, 2.

Hotta, L. K., & Tsay, R. S. (2012). Outliers in GARCH processes. In W. R. Bell, S. H. Holan, & T. S. McElroy (Eds.), *Economic time series: modeling and seasonality* (pp. 337–358). Boca Raton: Chapman & Hall.

Justel, A., Peña, D., & Tsay, R. S. (2001). Detection of outlier patches in autoregressive time series. *Statistica Sinica*, *11*, 651–673.

Kirkendall, N. J. (1992). Monitoring for outliers and level shifts in Kalman filter implementations of exponential smoothing. *Journal of Forecasting*, *11*, 543–560.

Koehler, A. B., Snyder, R. D., Ord, J. K., & Beaumont, A. (2012). A study of outliers in the exponential smoothing approach to forecasting. *International Journal of Forecasting*, *28*, 477–484.

Ljung, G. M. (1993). On outlier detection in time series. *Journal of the Royal Statistical Society. Series B. Methodological*, *55*, 559–567.

Luceño, A. (1998). Detecting possibly non-consecutive outliers in industrial time series. *Journal of the Royal Statistical Society. Series B. Methodological*, *60*, 295–310.

Maronna, R., Martin, R. D., & Yohai, V. (2006). *Robust statistics*. Chichester: Wiley.

Muler, N., Peña, D., & Yohai, V. (2009). Robust estimation for ARMA models. *The Annals of Statistics*, *37*, 816–840.

Muler, N., & Yohai, V. (2008). Robust estimates for GARCH models. *Journal of Statistical Planning and Inference*, *138*, 2918–2940.

Pankratz, A. (1993). Detecting and treating outliers in dynamic regression models. *Biometrika*, *80*, 847–854.

Penzer, J. (2007). State space models for time series with patches of unusual observations. *Journal of Time Series Analysis*, *28*, 629–645.

Peña, D., & Maravall, A. (1991). Interpolation, outliers and inverse autocorrelations. *Communications in Statistics. Theory and Methods*, *20*, 3175–3186.

Perron, P., & Rodríguez, G. (2003). Searching for additive outliers in nonstationary time series. *Journal of Time Series Analysis*, *24*, 193–220.

Sánchez, M. J., & Peña, D. (2010). The identification of multiple outliers in ARIMA models. *Communications in Statistics. Theory and Methods*, *32*, 1265–1287.

Tsay, R. S. (1986). Time series model specification in the presence of outliers. *Journal of the American Statistical Association*, *86*, 132–141.

Tsay, R. S. (1988). Outliers, level shifts and variance changes in time series. *Journal of Forecasting*, *7*, 1–20.

Tsay, R. S., Peña, D., & Pankratz, A. E. (2000). Outliers in multivariate time series. *Biometrika*, *87*, 789–804.

Part III
Complex Data Structures

Chapter 16
Qualitative Robustness of Bootstrap Approximations for Kernel Based Methods

Andreas Christmann, Matías Salibián-Barrera, and Stefan Van Aelst

16.1 Introduction

Current statistical applications are characterized by a wealth of large and complex data sets. There usually is a variable of main interest ("response" or "output values") and a number of potential explanatory measurements ("input values") that are to be used to either predict or describe the response. Our observations consist of n pairs $(x_1, y_1), \dots, (x_n, y_n)$, which will be assumed to be independent realizations of a random pair (X, Y). We are interested in obtaining a function $f : \mathcal{X} \to \mathcal{Y}$ such that $f(x)$ is a good predictor for the response y, if $X = x$ is observed. The prediction should be made in an automatic way. We refer to this process of determining a prediction method as "statistical machine learning", see, e.g., Vapnik (1995, 1998), Schölkopf and Smola (2002), Cucker and Zhou (2007), Smale and Zhou (2007). Here, by "good predictor" we mean that f minimizes the expected loss, i.e., the risk,

$$\mathcal{R}_{L,\mathrm{P}}(f) = \mathbb{E}_{\mathrm{P}}\big[L\big(X, Y, f(X)\big)\big],$$

where P denotes the unknown joint distribution of the random pair (X, Y) and $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, +\infty)$ is a fixed loss function. As a simple example, the least

A. Christmann (✉)
Department of Mathematics, University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany
e-mail: andreas.christmann@uni-bayreuth.de

M. Salibián-Barrera
Department of Statistics, University of British Columbia, Vancouver, Canada
e-mail: matias@stat.ubc.ca

S. Van Aelst
Department of Applied Mathematics and Computer Science, Ghent University, Ghent, Belgium
e-mail: Stefan.VanAelst@UGent.be

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_16, © Springer-Verlag Berlin Heidelberg 2013

squares loss $L(X, Y, f(X)) = (Y - f(X))^2$ yields the optimal predictor $f(x) = \mathbb{E}_{\mathrm{P}}(Y \mid X = x)$, $x \in \mathcal{X}$. Because P is unknown, we can neither compute nor minimize the risk $R_{L,\mathrm{P}}(f)$ directly.

Support Vector Machines provide a highly versatile framework to perform statistical machine learning in a wide variety of setups. Given a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ we consider predictors $f \in H$, where H denotes the corresponding reproducing kernel Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$. The space H includes, for example, all functions of the form $f(x) = \sum_{j=1}^{m} \alpha_j k(x, x_j)$ where x_j are arbitrary elements in $\mathcal{X}$ and $\alpha_j \in \mathbb{R}$, $1 \le j \le m$. To avoid overfitting, a support vector machine $f_{L,\mathrm{P},\lambda}$ is defined as the solution of a regularized risk minimization problem. More precisely,

$$f_{L,\mathrm{P},\lambda} = \arg \inf_{f \in H} \mathbb{E}_{\mathrm{P}} L\big(X, Y, f(X)\big) + \lambda \| f \|_H^2, \tag{16.1}$$

where $\lambda \in (0, \infty)$ is the regularization parameter. For a sample, $D = ((x_1, y_1), \dots, (x_n, y_n))$ the corresponding estimated function is given by

$$f_{L,\mathrm{D}_n,\lambda} = \arg \inf_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L\big(x_i, y_i, f(x_i)\big) + \lambda \| f \|_H^2, \tag{16.2}$$

where D_n denotes the empirical distribution based on D. Efficient algorithms to compute $\hat{f}_n := f_{L,\mathrm{D}_n,\lambda}$ exist for a number of different loss functions. Note that the optimization problem (16.2) corresponds to (16.1) when using D_n instead of P.

An important component of statistical analyses concerns quantifying and incorporating uncertainty (e.g., sampling variability) in the reported estimates. For example, one may want to include confidence bounds along the individual predicted values $\hat{f}_n(x_i)$ obtained from (16.2). Unfortunately, the sampling distribution of the estimated function $\hat{f}_n$ is typically unknown. Recently, Hable (2012) considered the asymptotic distribution of SVMs. Asymptotic confidence intervals based on those general results are always symmetric.

Here, we are interested in approximating the finite sample distribution of SVMs by Efron's bootstrap approach (Efron 1979), because confidence intervals based on the bootstrap approach can be asymmetric. The bootstrap provides an alternative way to estimate the sampling distribution of a wide variety of estimators. To fix ideas, consider a functional $S : \mathcal{M} \to \mathcal{W}$, where $\mathcal{M}$ is a set of probability measures and $\mathcal{W}$ denotes a metric space. Many estimators can be included in this framework. Simple examples include the sample mean (with functional $S(\mathrm{P}) = \int Z \, d\mathrm{P}$) and M-estimators (with functional defined implicitly as the solution to the equation $\mathbb{E}_{\mathrm{P}} \Psi(Z, S(\mathrm{P})) = 0$). Let $\mathcal{B}(\mathcal{Z})$ be the Borel σ-algebra on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and denote the set of all Borel probability measures on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ by $\mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$. Then, it follows that (16.1) defines an operator $S : \mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z})) \to H$ whose value is given by $S(P) = f_{L,\mathrm{P},\lambda}$, i.e., the support vector machine. Moreover, the estimator in (16.2) satisfies

$$f_{L,\mathrm{D}_n,\lambda} = S(\mathrm{D}_n),$$

where

$$\mathrm{D}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(x_i,y_i)}$$

is the empirical distribution based on the sample $D = ((x_1, y_1), \dots, (x_n, y_n))$ and $\delta_{(x_i,y_i)}$ denotes the Dirac measure at the point (x_i, y_i).

More generally, let $Z_i = (X_i, Y_i)$, $i = 1, \dots, n$, be independent and identically distributed (i.i.d.) random variables with distribution P, and let

$$S_n(Z_1, \dots, Z_n) = S(\mathrm{P}_n)$$

be the corresponding estimator, where

$$\mathrm{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}.$$

Denote the distribution of $S(\mathrm{P}_n)$ by $\mathscr{L}_n(S; \mathrm{P}) = \mathscr{L}(S(\mathrm{P}_n))$. If P was known to us, we could estimate this sampling distribution by drawing a large number of random samples from P and evaluating our estimator on them. The basic idea behind the bootstrap is to replace the unknown distribution P by an estimate $\hat{\mathrm{P}}$. Here we will consider the natural non-parametric estimator given by the sample empirical distribution P_n. In other words, we estimate the distribution of our estimator of interest by its sampling distribution when the data are generated by P_n. In symbols, the bootstrap proposes to use $\widehat{\mathcal{L}_n(S; \mathrm{P})} = \mathscr{L}_n(S; \mathrm{P}_n)$. Since this distribution is generally unknown, in practice one uses Monte Carlo simulation to estimate it by repeatedly evaluating the estimator on samples drawn from D_n. Note that drawing a sample from D_n means that n observations are drawn *with replacement* from the original n observations $(x_1, y_1), \dots, (x_n, y_n)$.

It is easy to see that even a small proportion of outliers in the sample might have an important effect on the bootstrap samples drawn from D_n. In particular, due to the sampling with replacement the number of outliers in the bootstrap samples might be much higher than that present in the original sample, and thus seriously affect the estimated distribution of the estimator.

The main goal of this contribution is to show that bootstrap approximations of an estimator which is based on a continuous operator S from the set of Borel probability distributions defined on a compact metric space into a complete separable metric space is stable in the sense of qualitative robustness. Intuitively, this means that small deviations from the distribution P that generates the data can only produce small perturbations in the bootstrap estimate of the sampling distribution of the estimator. In other words, the bootstrap distribution of these estimators will not change much if the data contains a relatively small proportion of outliers or not.

As a special case, we will show that bootstrap approximations for many (general) support vector machines are qualitatively robust, both for the real-valued SVM risk functional and for the H-valued SVM operator itself. Our results only require conditions on the loss function and on the kernel, but not on the unknown distribution P. Hence, these conditions can be checked completely in advance and are not data dependent. Our work generalizes previous results in Cuevas and Romo (1993).

The rest of this chapter has the following structure. Section 16.2 lists some tools which we need to prove our general results in Sect. 16.3. Section 16.4 contains the results on bootstrap approximations of SVMs. A short discussion is given in Sect. 16.5 where we also mention some related work.

16.2 Some Tools

For the proofs of our results in Sect. 16.3, we need Theorems 16.1 and 16.2 listed below.

To state Theorem 16.1 on uniform Glivenko–Cantelli classes, we need the following notation. For any metric space $(\mathcal{S}, d)$ and real-valued function $f : \mathcal{S} \to \mathbb{R}$, we denote the *bounded Lipschitz norm* of f by

$$\|f\|_{\mathrm{BL}} := \sup_{x\in\mathcal{S}} \left|f(x)\right| + \sup_{x,y\in\mathcal{S},x\neq y} \frac{|f(x) - f(y)|}{d(x, y)}. \tag{16.3}$$

Let $\tilde{F}$ be a set of measurable functions from $(\mathcal{S}, \mathcal{B}(\mathcal{S})) \to (\mathbb{R}, \mathcal{B})$. For any function $G : \tilde{F} \to \mathbb{R}$ (such as a signed measure) define

$$\|G\|_{\tilde{F}} := \sup\left\{\left|G(f)\right| : f \in \tilde{F}\right\}. \tag{16.4}$$

The next result uses outer probabilities, we refer to van der Vaart and Wellner (1996) for details.

Theorem 16.1 (Dudley et al. 1991, Proposition 12) *For any separable metric space* $(\mathcal{S}, d)$ *and* $M \in (0, \infty)$,

$$\tilde{\mathcal{F}}_M := \left\{f : \left(\mathcal{S}, \mathcal{B}(\mathcal{S})\right) \to (\mathbb{R}, \mathcal{B});\ \|f\|_{\mathrm{BL}} \leq M\right\} \tag{16.5}$$

is a universal Glivenko–Cantelli class. It is a uniform Glivenko–Cantelli class, i.e., for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \sup_{\nu\in\mathcal{M}_1(\mathcal{S},\mathcal{B}(\mathcal{S}))} \Pr{}^*\left(\sup_{m\geq n} \|\nu_m - \nu\|_{\tilde{\mathcal{F}}_M} > \varepsilon\right) = 0, \tag{16.6}$$

if and only if $(\mathcal{S}, d)$ *is totally bounded. Here,* $\Pr^*$ *denotes the outer probability.*

Note that the term $\|\nu_m - \nu\|_{\tilde{\mathcal{F}}_M}$ in (16.6) equals the *bounded Lipschitz metric* d_{BL} of the probability measures ν_m and ν if $M = 1$, i.e.,

$$\begin{aligned}\|\nu_m - \nu\|_{\tilde{\mathcal{F}}_1} &= \sup_{f \in \tilde{\mathcal{F}}_1} \left|(\nu_m - \nu)(f)\right| \\ &= \sup_{f;\, \|f\|_{\mathrm{BL}} \le 1} \left| \int f \, d\nu_m - \int f \, d\nu \right| =: d_{\mathrm{BL}}(\nu_m, \nu), \qquad (16.7)\end{aligned}$$

see Dudley (2002, p. 394). Hence, Theorem 16.1 can be interpreted as a generalization of Cuevas and Romo (1993, Lemma 1, p. 186), which states that if $A \subset \mathbb{R}$ is a finite interval, then $d_{\mathrm{BL}}(\mathrm{P}_m, \mathrm{P})$ converges almost surely to 0 uniformly in $\mathrm{P} \in \mathcal{M}_1(A, \mathcal{B}(A))$. For various characterizations of Glivenko–Cantelli classes, we refer to Talagrand (1987, Theorem 22), Ledoux and Talagrand (1991), and Dudley (1999).

We next list the other main result we need for the proof of Theorem 16.4. This result is an analogon of the famous Strassen theorem, but for the bounded Lipschitz metric d_{BL} instead of for the Prohorov metric.

Theorem 16.2 (Huber 1981, Theorem 4.2, p. 30) *Let $\mathcal{Z}$ be a Polish space with topology $\tau_{\mathcal{Z}}$. Let d_{BL} be the bounded Lipschitz metric defined on the set $\mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ of all Borel probability measures on $\mathcal{Z}$. Then the following two statements are equivalent*:

(i) *There are random variables ξ_1 with distribution ν_1 and ξ_2 with distribution ν_2 such that $\mathbb{E}[d_{\mathrm{BL}}(\xi_1, \xi_2)] \le \varepsilon$.*
(ii) *$d_{\mathrm{BL}}(\nu_1, \nu_2) \le \varepsilon$.*

16.3 On Qualitative Robustness of Bootstrap Estimators

Unless otherwise mentioned, we will use the Borel σ-algebra $\mathcal{B}(\mathcal{Z})$ on a metric space $(\mathcal{Z}, d_{\mathcal{Z}})$ and denote the Borel σ-algebra on $\mathbb{R}$ by $\mathcal{B}$.

Assumption 16.1 Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, where μ is unknown, $(\mathcal{Z}, d_{\mathcal{Z}})$ be a compact metric space, and $\mathcal{B}(\mathcal{Z})$ be the Borel σ-algebra on $\mathcal{Z}$. Denote the set of all Borel probability measures on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ by $\mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$. On $\mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ we use the Borel σ-algebra $\mathcal{B}(\mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z})))$ and the bounded Lipschitz metric d_{BL} defined by (16.7). Let S be a statistical operator defined on the set $\mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ with values in a complete, separable metric space $(\mathcal{W}, d_{\mathcal{W}})$ enclipped with its Borel σ-algebra $\mathcal{B}(\mathcal{W})$. Let $Z, Z_n : (\Omega, \mathcal{A}, \mu) \to (\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$, $n \in \mathbb{N}$, be independent and identically distributed random variables and denote the image measure by $\mathrm{P} := Z \circ \mu$. Let $S_n(Z_1, \dots, Z_n)$ be a statistic with values in $(\mathcal{W}, \mathcal{B}(\mathcal{W}))$. Denote the empirical measure of $(Z_1, \dots, Z_n)$ by $\mathrm{P}_n := \frac{1}{n} \sum_{i=1}^n \delta_{Z_i}$. The statistic S_n is defined via the operator

$$S : \big(\mathcal{M}_1\big(\mathcal{Z}, \mathcal{B}(\mathcal{Z})\big), \mathcal{B}\big(\mathcal{M}_1\big(\mathcal{Z}, \mathcal{B}(\mathcal{Z})\big)\big)\big) \to \big(\mathcal{W}, \mathcal{B}(\mathcal{W})\big)$$

where $S(\mathrm{P}_n) = S_n(Z_1, \ldots, Z_n)$. Denote the distribution of $S_n(Z_1, \ldots, Z_n)$ when $Z_i \overset{\text{i.i.d.}}{\sim} \mathrm{P}$ by $\mathscr{L}_n(S; \mathrm{P}) := \mathscr{L}(S_n(Z_1, \ldots, Z_n))$. Accordingly, we denote the distribution of $S_n(Z_1, \ldots, Z_n)$ when $Z_i \overset{\text{i.i.d.}}{\sim} \mathrm{P}_n$ by $\mathscr{L}_n(S; \mathrm{P}_n)$.

We are interested in estimating the sampling distribution $\mathscr{L}_n(S; \mathrm{P})$ of the statistic $S_n(Z_1, \ldots, Z_n)$. Efron (1979, 1982) proposed to use the bootstrap for this purpose. The intuitive idea can be expressed as follows. If $Z_i \overset{\text{i.i.d.}}{\sim} \mathrm{P}$ and P was known, we could calculate (or simulate) the distribution $\mathscr{L}_n(S; \mathrm{P})$ of $S_n(Z_1, \ldots, Z_n)$. Since P is unknown, the bootstrap proposes to estimate it by the empirical distribution P_n of our sample. Formally, the bootstrap estimates the unknown distribution $\mathscr{L}_n(S; \mathrm{P})$ with $\mathscr{L}_n(S; \mathrm{P}_n)$. Since the latter is also generally unknown, in practice computer simulations are used to estimate $\mathscr{L}_n(S; \mathrm{P}_n)$. This corresponds to generating many samples $Z_1^*, \ldots, Z_n^*$, with $Z_i^* \overset{\text{i.i.d.}}{\sim} \mathrm{P}_n$ and computing the corresponding estimates $S_n(Z_1^*, Z_2^*, \ldots, Z_n^*)$. Note that the bootstrap approximations $\mathscr{L}_n(S; \mathrm{P}_n)$ are probability measure-valued random variables with values in $\mathcal{M}_1(\mathcal{W}, \mathcal{B}(\mathcal{W}))$.

Following Cuevas and Romo (1993), we call a sequence of bootstrap approximations $\mathscr{L}_n(S; \mathrm{P}_n)$ *qualitatively robust* at $\mathrm{P} \in \mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ if the sequence of transformations $(g_n)_{n\in\mathbb{N}}$ defined by

$$g_n(\mathrm{Q}) = \mathscr{L}\big(\mathscr{L}_n(S; \mathrm{Q}_n)\big), \quad \mathrm{Q} \in \mathcal{M}_1\big(\mathcal{Z}, \mathcal{B}(\mathcal{Z})\big),$$

is asymptotically equicontinuous at $\mathrm{P} \in \mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$, i.e. if

$$\begin{aligned}&\forall \varepsilon > 0\ \exists \delta > 0\ \exists n_0 \in \mathbb{N}:\\ &\quad d_{\mathrm{BL}}(\mathrm{Q}, \mathrm{P}) < \delta \Rightarrow \sup_{n \ge n_0} d_{\mathrm{BL}}\Big(\mathscr{L}\big(\mathscr{L}_n(S; \mathrm{Q}_n)\big), \mathscr{L}\big(\mathscr{L}_n(S; \mathrm{P}_n)\big)\Big) < \varepsilon. \quad (16.8)\end{aligned}$$

The above definition can be interpreted as follows. If the measures P and Q are close, then the distribution of the bootstrap estimators when samples are drawn from P and Q also remain close. In other words, small deviations in the distributions generating the data only produce mild differences in the distribution of the bootstrap estimators.

As in Cuevas and Romo (1993), we call a sequence of statistics $(S_n)_{n\in\mathbb{N}}$ *uniformly qualitatively robust in a neighborhood* $\mathcal{U}(\mathrm{P}_0)$ *of* $\mathrm{P}_0 \in \mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ if

$$\begin{gathered}\exists n_0 \in \mathbb{N} \quad \forall \varepsilon > 0 \quad \forall n \ge n_0\ \exists \delta > 0 \quad \forall \mathrm{P} \in \mathcal{U}(\mathrm{P}_0):\\ d_{\mathrm{BL}}(\mathrm{Q}, \mathrm{P}) < \delta \Rightarrow d_{\mathrm{BL}}\big(\mathscr{L}_n(S; \mathrm{Q}), \mathscr{L}_n(S; \mathrm{P})\big) < \varepsilon.\end{gathered}$$

The following two results and Theorem 16.5 in the next section constitute the core of this paper. They show that bootstrap approximations of an estimator which is based on a continuous operator S from the set of Borel probability distributions defined on a compact metric space into a complete separable metric space is stable in the sense of the qualitative robustness as defined in (16.8).

Theorem 16.3 *If Assumption 16.1 is valid and if S is uniformly continuous in a neighborhood $\mathcal{U}(\mathrm{P}_0)$ of $\mathrm{P}_0 \in \mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$, then the sequence $(S_n(Z_1, \ldots, Z_n))_{n\in\mathbb{N}}$ is uniformly qualitatively robust in $\mathcal{U}(\mathrm{P}_0)$.*

Proof We closely follow the proof by Cuevas and Romo (1993, Theorem 2). However, we use Theorem 16.1 instead of their Lemma 1 and we use Cuevas (1988, Lemma 1) instead of Hampel (1971, Lemma 1). Let $\mathcal{P}_n \subset \mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ be the set of empirical distributions of order $n \in \mathbb{N}$, i.e.

$$\mathcal{P}_n := \left\{ \mathrm{P}_n \in \mathcal{M}_1\big(\mathcal{Z}, \mathcal{B}(\mathcal{Z})\big);\, \exists (z_1, \ldots, z_n) \in \mathcal{Z}^n \text{ such that } \mathrm{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{z_i} \right\},$$

and let $\mathcal{E}_n \subset \mathcal{P}_n$. If misunderstandings are unlikely, we identify $\mathcal{E}_n$ with the set $\{z_1, \ldots, z_n\}$ of atoms.

It is enough to show that

$$\forall \varepsilon > 0\ \exists \delta > 0 \quad \forall \mathrm{P} \in \mathcal{U}(\mathrm{P}_0)\ \exists \text{ sequence } (\mathcal{E}_n)_{n\in\mathbb{N}} \subset \mathcal{P}_n$$

such that $\mathrm{P}^n(\mathcal{E}_n) > 1 - \varepsilon$ and for all $\mathrm{Q}_n \in \mathcal{E}_n$ and for all $\tilde{\mathrm{Q}}_n \in \mathcal{P}_n$ we have

$$d_{\mathrm{BL}}(\mathrm{Q}_n, \tilde{\mathrm{Q}}_n) < \delta \Rightarrow d_{\mathcal{W}}\big(S(\mathrm{Q}_n), S(\tilde{\mathrm{Q}}_n)\big) < \varepsilon.$$

From this, we obtain that $(S_n)_{n\in\mathbb{N}}$ is uniformly qualitatively robust by Cuevas (1988, Lemma 1).

Let $\varepsilon > 0$. Since the operator S is uniformly continuous in $\mathcal{U}(\mathrm{P}_0)$ we obtain

$$\exists \delta_0 > 0 \quad \forall \mathrm{P} \in \mathcal{U}(\mathrm{P}_0) : d_{\mathrm{BL}}(\mathrm{P}, \mathrm{Q}) < \delta_0 \Rightarrow d_{\mathcal{W}}\big(S(\mathrm{P}), S(\mathrm{Q})\big) < \varepsilon/2. \tag{16.9}$$

Hence by Theorem 16.1 for the special case $M = 1$ and by (16.7), we get

$$\exists n_0 \in \mathbb{N} : \sup_{\mathrm{P}\in\mathcal{U}(\mathrm{P}_0)} \mathrm{Pr}^*\Big(\sup_{n\geq n_0} d_{\mathrm{BL}}(\mathrm{P}_n, \mathrm{P}) < \delta_0\Big) > 1 - \varepsilon.$$

For $n \geq n_0$ and $\mathrm{P} \in \mathcal{U}(\mathrm{P}_0)$, define

$$\mathcal{E}_{n,\mathrm{P}} := \big\{\mathrm{Q}_n \in \mathcal{P}_n : d_{\mathrm{BL}}(\mathrm{Q}_n, \mathrm{P}) < \delta_0/2\big\}.$$

It follows, that $\mathrm{P}^n(\mathcal{E}_{n,\mathrm{P}}) > 1 - \varepsilon$ together with $\mathrm{Q}_n \in \mathcal{E}_{n,\mathrm{P}}$ and $d_{\mathrm{BL}}(\mathrm{Q}_n, \tilde{\mathrm{Q}}_n) < \delta_0/2$ implies that

$$d_{\mathrm{BL}}(\mathrm{Q}_n, \mathrm{P}) < \delta_0/2 \quad \text{and} \quad d_{\mathrm{BL}}(\tilde{\mathrm{Q}}_n, \mathrm{P}) < \delta_0.$$

The triangle inequality thus yields due to (16.9)

$$d_{\mathcal{W}}\big(S(\mathrm{Q}_n), S(\tilde{\mathrm{Q}}_n)\big) \leq d_{\mathcal{W}}\big(S(\mathrm{Q}_n), S(\mathrm{P})\big) + d_{\mathcal{W}}\big(S(\mathrm{P}), S(\tilde{\mathrm{Q}}_n)\big) < \varepsilon,$$

from which the assertion follows. □

Theorem 16.4 *If Assumption 16.1 is valid and if the sequence* $(S_n(Z_1, \ldots, Z_n))_{n\in\mathbb{N}}$ *is uniformly qualitatively robust in a neighborhood* $\mathcal{U}(\mathrm{P}_0)$ *of* $\mathrm{P}_0 \in \mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$, *then the sequence* $\mathscr{L}_n(S; \mathrm{P}_n)$ *of bootstrap approximations of* $\mathscr{L}_n(S; \mathrm{P})$ *is qualitatively robust for* P_0.

Proof The proof mimics the proof of Cuevas and Romo (1993, Theorem 3), but uses Theorem 16.1 instead of Cuevas and Romo (1993, Lemma 1).

Fix $\mathrm{P}_0 \in \mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ and $\varepsilon > 0$. By the uniform qualitative robustness of $(S_n)_{n\in\mathbb{N}}$ in $\mathcal{U}(\mathrm{P}_0)$, there exists $n_0 \in \mathbb{N}$ such that for all $\varepsilon > 0$ there exists $\delta > 0$ such that

$$d_{\mathrm{BL}}(\mathrm{Q}, \mathrm{P}) < \delta \Rightarrow \sup_{m\geq n_0} \sup_{\mathrm{P}\in\mathcal{U}(\mathrm{P}_0)} d_{\mathrm{BL}}\big(\mathscr{L}_m(S; \mathrm{Q}), \mathscr{L}_m(S; \mathrm{P})\big) < \varepsilon. \tag{16.10}$$

Define $\delta_1 := \delta/2$. Due to Theorem 16.1 for the special case $M = 1$ and by (16.7), we have, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \sup_{\mathrm{P}\in\mathcal{M}_1(\mathcal{Z},\mathcal{B}(\mathcal{Z}))} \mathrm{Pr}^*\Big(\sup_{m\geq n} d_{\mathrm{BL}}(\mathrm{P}_m, \mathrm{P}) > \varepsilon\Big) = 0.$$

Hence (16.10) and Varadarajan's theorem on the almost sure convergence of empirical measures to a Borel probability measure defined on a separable metric space, see, e.g., Dudley (2002, Theorem 11.4.1, p. 399), yields for the empirical distributions Q_n from Q and $\mathrm{P}_{0,n}$ from P_0 that,

$$\begin{aligned} &\exists n_1 > n_0 \quad \forall n \geq n_1 : \\ &\quad d_{\mathrm{BL}}(\mathrm{Q}, \mathrm{P}_0) < \delta_1 \Rightarrow d_{\mathrm{BL}}(\mathrm{Q}_n, \mathrm{P}_{0,n}) < \delta \quad \text{almost surely.} \end{aligned} \tag{16.11}$$

It follows from the uniform qualitative robustness of $(S_n)_{n\in\mathbb{N}}$, see (16.10), that

$$\begin{aligned} &\exists n_1 \in \mathbb{N} \quad \forall \varepsilon > 0 \quad \forall n \geq n_1 \; \exists \delta > 0 \quad \forall \mathrm{P} \in \mathcal{U}(\mathrm{P}_0) : \\ &\quad d_{\mathrm{BL}}(\mathrm{Q}, \mathrm{P}) < \delta \Rightarrow d_{\mathrm{BL}}\big(\mathscr{L}_n(S; \mathrm{Q}_n), \mathscr{L}_n(S; \mathrm{P}_{0,n})\big) < \varepsilon \text{ almost surely.} \end{aligned} \tag{16.12}$$

For notational convenience, we write for the sequences of bootstrap estimators

$$\xi_{1,n} := \mathscr{L}_n(S; \mathrm{Q}_n), \qquad \xi_{2,n} := \mathscr{L}_n(S; \mathrm{P}_{0,n}), \quad n \in \mathbb{N}.$$

Note that $\xi_{1,n}$ and $\xi_{2,n}$ are (measure-valued) random variables with values in the set $\mathcal{M}_1(\mathcal{W}, \mathcal{B}(\mathcal{W}))$. We denote the distribution of $\xi_{j,n}$ by $\mu_{j,n}$ for $j \in \{1, 2\}$ and $n \in \mathbb{N}$. Hence, (16.12) yields

$$d_{\mathrm{BL}}(\xi_{1,n}, \xi_{2,n}) < \varepsilon \text{ almost surely for all } n \geq n_1$$

and it follows

$$\mathbb{E}\big[d_{\mathrm{BL}}(\xi_{1,n}, \xi_{2,n})\big] \leq \varepsilon \quad \forall n \geq n_1.$$

Now an application of an analogon of Strassen's theorem, see Theorem 16.2, yields

$$\sup_{n\geq n_1} d_{\mathrm{BL}}\big(\mathscr{L}(\xi_{1,n}), \mathscr{L}(\xi_{2,n})\big) \leq \varepsilon,$$

which completes the proof, because

$$\mathscr{L}(\xi_{1,n}) = \mathscr{L}\big(\mathscr{L}_n(S; \mathrm{Q}_n)\big) \quad \text{and} \quad \mathscr{L}(\xi_{2,n}) = \mathscr{L}\big(\mathscr{L}_n(S; \mathrm{P}_{0,n})\big). \qquad \square$$

As an immediate consequence of these results we obtain the following corollary.

Corollary 16.1 *If Assumption 16.1 is valid and if S is a continuous operator, then the sequence $\mathscr{L}_n(S; \mathrm{P}_n)$, $n \in \mathbb{N}$, of bootstrap approximations of $\mathscr{L}_n(S; \mathrm{P})$ is qualitatively robust for all* $\mathrm{P} \in \mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$.

Remark 16.1 The Theorems 16.3 and 16.4 can be interpreted as a generalization of Theorems 2 and 3 in Cuevas and Romo (1993), who considered, for some fixed finite interval $A \subset \mathbb{R}$, the case of a statistical functional S defined on $\mathcal{M}_1(A, \mathcal{B}(A))$, i.e., $\mathcal{Z} = A \subset \mathbb{R}$ a finite interval and $\mathcal{W} = \mathbb{R}$. We considered the case, that $\mathcal{Z}$ is a compact metric space and $\mathcal{W}$ is a complete separable metric space.

16.4 On Qualitative Robustness of Bootstrap SVMs

In this section, we will apply the previous results to support vector machines (SVMs) based on some general loss function L and some reproducing kernel Hilbert space (RKHS) H with kernel k. Such SVMs belong to the modern class of statistical machine learning methods based on kernels. In other words, we will consider the special case when $\mathcal{W}$ is the RKHS used by a SVM. Note that often rich RKHSs with an infinite dimension are used in machine learning theory to make it possible that SVMs based on the minimization over such function spaces are universally consistent.

Recall that SVMs are defined by a minimization problem as in (16.1). In what follows, we will consider a fixed loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, +\infty)$, regularizing constant $\lambda \in (0, \infty)$, and kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (with associated RKSH H). Such a definition generally implies a moment condition on the distribution of Y given $X = x$. To weaken this type of assumption, we use a "shifted loss" $L^\star(x, y, t) := L(x, y, t) - L(x, y, 0)$, $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$. This idea was already used by Huber (1967) in the context of M-estimators. Note that SVMs can be considered as a generalization of M-estimators in the sense that SVMs are H-valued statistics defined as solutions of a minimization problem with an additional regularization term. The use of shifted loss functions enables us to define SVMs, which are based on a Lipschitz continuous convex loss function and on a bounded continuous kernel, on the whole space of probability measures and eliminates classical moment conditions on the conditional distribution of Y given $X = x$. This is in general not true for the non Lipschitz continuous least squares loss function, if $\mathcal{Y} \subset \mathbb{R}$ is unbounded. Since $L^\star$, k (or H), and λ are fixed, to simplify our notation, we will denote the SVM operator by S and the corresponding risk by R instead of $S_{L^\star,H,\lambda}$ and $R_{L^\star,H,\lambda}$ in the next definition, respectively.

Definition 16.1 The *SVM operator* $S : \mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z})) \to H$ is defined by

$$S(\mathrm{P}) := f_{L^\star,\mathrm{P},\lambda} := \arg\min_{f \in H} \mathbb{E}_{\mathrm{P}} L^\star\big(X, Y, f(X)\big) + \lambda \|f\|_H^2. \tag{16.13}$$

The *SVM risk functional* $R : \mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z})) \to \mathbb{R}$ is defined by

$$R(\mathrm{P}) := \mathbb{E}_{\mathrm{P}} L^{\star}\big(X, Y, S(\mathrm{P})(X)\big) = \mathbb{E}_{\mathrm{P}} L^{\star}\big(X, Y, f_{L^{\star},\mathrm{P},\lambda}(X)\big). \tag{16.14}$$

Our stability result for the bootstrap approximations of SVMs (see Theorem 16.5 below) requires the following assumptions on the loss and kernel functions.

Assumption 16.2 Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be a compact metric space with metric $d_{\mathcal{Z}}$, where $\mathcal{Y} \subset \mathbb{R}$. Let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function such that L is continuous and convex with respect to its third argument and that L is uniformly Lipschitz continuous with respect to its third argument with uniform Lipschitz constant $|L|_1 \in (0, \infty)$, i.e. $|L|_1$ is the smallest constant c such that $\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |L(x, y, t) - L(x, y, t')| \leq c|t - t'|$ for all $t, t' \in \mathbb{R}$. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous kernel with reproducing kernel Hilbert space H and assume that k is bounded by $\|k\|_\infty := (\sup_{x\in\mathcal{X}} k(x, x))^{1/2} \in (0, \infty)$. Let $\lambda \in (0, \infty)$.

Under Assumption 16.2 the SVM operator S in (16.13) is well defined because $S(\mathrm{P}) \in H$ exists and is unique, the SVM risk functional R in (16.14) is also well defined because $R(\mathrm{P}) \in \mathbb{R}$ exists and is unique, and it holds that for all probability measures P on $\mathcal{X} \times \mathcal{Y}$ that

$$\big\|S(\mathrm{P})\big\|_\infty \leq \frac{1}{\lambda}|L|_1 \|k\|_\infty^2 < \infty \quad \text{and} \quad \big|R(\mathrm{P})\big| \leq \frac{1}{\lambda}|L|_1^2 \|k\|_\infty^2 < \infty, \tag{16.15}$$

see Christmann et al. (2009, Theorem 5, Theorem 6, (17), (18)).

Note that the conditions on L and k in Assumption 16.2 do *not* depend on the unknown distribution P or on the data set d. Hence, they are easy to verify, and can be considered as standard assumptions for statistically robust SVMs, see, e.g., Christmann and Steinwart (2007), Steinwart and Christmann (2008, Chap. 10), Christmann et al. (2009), and Hable and Christmann (2011). However, the assumption that $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is a *compact* metric space is more restrictive than the assumption that $\mathcal{X}$ is a Polish space, which is by now often assumed to prove consistency or learning rates of SVMs. Our proof relies on the compactness assumption, but it would be interesting to investigate whether this assumption can be weakend, e.g., to σ-compactness which would cover the special case of the standard Euclidean space $\mathcal{Z} = \mathbb{R}^d \times \mathbb{R}$.

We can now state our main result on the robustness of the bootstrap approach for support vector machines.

Theorem 16.5 *If the general Assumption* 16.1 *and Assumption* 16.2 *are valid, then the SVM operator S and the SVM risk functional R fulfill*:

(i) *The sequence $\mathscr{L}_n(S; \mathrm{P}_n)$, $n \in \mathbb{N}$, of bootstrap SVM estimators of $\mathscr{L}_n(S; \mathrm{P})$ is qualitatively robust for all* $\mathrm{P} \in \mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$.

(ii) *The sequence $\mathscr{L}_n(R; \mathrm{P}_n)$, $n \in \mathbb{N}$, of bootstrap SVM risk estimators of $\mathscr{L}_n(R; \mathrm{P})$ is qualitatively robust for all* $\mathrm{P} \in \mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$.

Proof See Christmann et al. (2011, Theorem 8). The proof is somewhat technical and consists in checking that our Theorems 16.3 and 16.4 are applicable. □

The above theorem states that for SVMs satisfying Assumptions 16.1 and 16.2, the bootstrap estimators (for the risk and for the SVM itself) of their sampling distributions are resistant to small perturbations in the data-generating process. In other words, when datasets contain a small proportion of outliers, the bootstrap estimators do not deviate much from those that would have been obtained with a clean data set. Outliers in the sample thus do not have a large effect on the distribution of the bootstrap estimators.

We now list some loss functions L and kernels k which have the properties described in Assumption 16.2. Therefore, the sequence of bootstrap approximations of SVMs based on any combination of these loss and kernel functions are qualitatively robust due to Theorem 16.5.

Example 16.1 The assumptions on the loss function L from Assumption 16.2 are satisfied for example in the following cases.

(i) for classification with $\mathcal{Y} := \{-1, +1\}$:

- *hinge loss*: $L(x, y, t) := \max\{0, 1 - yt\}$,
- *logistic loss*: $L(x, y, t) := \ln(1 + \exp(-yt))$,

(ii) for regression with $\mathcal{Y} \subset \mathbb{R}$:

- *ϵ-insensitive loss* for some $\epsilon \in (0, \infty)$: $L(x, y, t) := \max\{0, |y - t| - \epsilon\}$,
- *$L1$-loss*: $L(x, y, t) := |y - t|$,
- *Huber's loss* for some $c > 0$:

$$L(x, y, t) := \begin{cases} 0.5(y-t)^2, & \text{if } |y-t| \le c, \\ c|y-t| - 0.5c^2, & \text{if } |y-t| > c, \end{cases}$$

- *logistic loss* for some $\gamma \in (0, \infty)$:

$$L(x, y, t) := -\gamma \ln \frac{4e^r}{(1+e^r)^2}, \quad \text{where } r := \frac{y-t}{\gamma},$$

(iii) for τ-quantile regression with $\mathcal{Y} \subset \mathbb{R}$ for level $\tau \in (0, 1)$:

- pinball loss:

$$L(x, y, t) := \begin{cases} (\tau - 1)(y-t), & \text{if } y - t < 0, \\ \tau(y-t), & \text{if } y - t \ge 0. \end{cases}$$

We mention that the classical least squares loss function defined by $L(x, y, t) := (y - t)^2$ for $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$ is of course locally Lipschitz continuous but *not* Lipschitz continuous if $\mathcal{Y} = \mathbb{R}$. SVMs based on this particular loss function are sometimes called least squares SVMs, see, e.g., Suykens et al. (2002), and can be considered as a special case of regularized least squares regression. It is well-known that the use of the least squares loss function in regression purposes with $\mathcal{Y} = \mathbb{R}$ yields SVMs which are *not* robust with respect to several common notions of statistical robustness even if the kernel is bounded.

Example 16.2 The assumptions on the kernel k from Assumption 16.2 are satisfied for example in the following cases.

(i) Gaussian radial basis function (RBF) kernels (for some $\gamma > 0$):

$$k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \quad k\big(x, x'\big) := \exp\big(-\gamma \left\| x - x' \right\|_2^2\big).$$

This kernel is interesting for machine learning purposes because its RKHS is dense in $L_p(\mu)$ for any finite measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and any $p \in [1, \infty)$. Furthermore, if we restrict k to $\tilde{k} := k_{|\mathcal{X}\times\mathcal{X}}$, where $\mathcal{X} \subset \mathbb{R}^d$ is compact, then the RKHS $\tilde{H}$ corresponding to the kernel $\tilde{k}$ is dense in $\mathcal{C}(\mathcal{X})$, and thus $\tilde{k}$ is a *universal* kernel. For more properties of this kernel and its RKHS we refer, e.g., to Steinwart and Christmann (2008, Chaps. 4.4 and 4.6).

(ii) Wendland RBF kernels (for some $d \in \mathbb{N}$ and $\ell \in \mathbb{N} \cup \{0\}$):

$$k_{d,\ell} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \quad k_{d,\ell}\big(x, x'\big) := \begin{cases} p_{d,\ell}(\|x - x'\|_2), & \text{if } \|x - x'\|_2 \in [0, 1], \\ 0, & \text{else}, \end{cases}$$

where $p_{d,\ell}$ is a certain univariate polynomial of degree $\lfloor \frac{d}{2} \rfloor + 3\ell + 1$, with $d \geq 3$ if $\ell = 0$. Wendland RBF kernels have a *bounded support*, which is not true for Gaussian RBF kernels. Wendland kernels are interesting for machine learning purposes because the native space of the basis functions is the classical *Sobolev space* $H^{d/2+\ell+1/2}(\mathbb{R}^d)$, see Wendland (2005, Theorems 9.13, 10.35) for details.

(iii) The Laplacian RBF kernels (for some $\gamma > 0$)

$$k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \quad k\big(x, x'\big) := \exp\big(-\gamma \left\| x - x' \right\|_2\big)$$

and the related RBF kernels (for some $\gamma > 0$):

$$k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \quad k\big(x, x'\big) := \exp\big(-\gamma \left\| x - x' \right\|_1\big).$$

These two classes of kernels are interesting for machine learning purposes because their RKHSs H are *completely separating* in the sense of De Vito et al. (2012), which is not true for the Gaussian RBF kernel, see De Vito et al. (2012, p. 13). An RKHS H separates a subset $C \subset \mathcal{X}$, if, for all $x \notin C$, there exists some $f \in H$ such that

$$f(x) \neq 0 \quad \text{and} \quad f(y) = 0 \quad \forall y \in C.$$

An RKHS H with kernel k, which satisfies $k(\cdot, x) \neq k(\cdot, x')$ for all $x \neq x'$ is called completely separating, if H separates all the subsets $C \subset \mathcal{X}$ which are closed with respect to the metric d_k defined by

$$\begin{aligned} d_k(x, x') &:= \| k(\cdot, x) - k(\cdot, x') \|_H \\ &= \sqrt{k(x, x) + k(x', x') - 2k(x, x')}, \quad x, x' \in \mathcal{X}. \end{aligned}$$

We refer to Scovel et al. (2010) for a large class of bounded RBF kernels and their corresponding RKHSs.

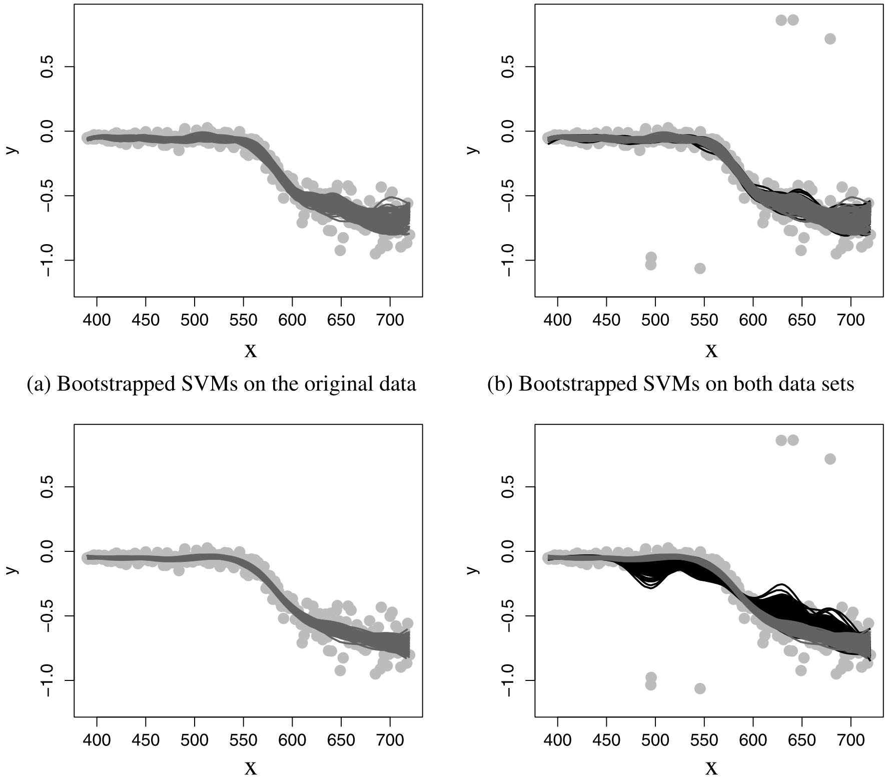
Example 16.3 To illustrate the practical implications of our results, consider the LIDAR data of Ruppert et al. (2003), available in package `SemiPar` for `R` (R Core Team 2012). These data consist of 221 observations from a light detection and ranging (LIDAR) experiment. The response variable is the logarithm of the ratio of the received light from two laser sources and the explanatory variable is the distance travelled by the light before it was reflected back. We used a grid search to determine appropriate hyperparameters. We consider an SVM with an ϵ-insensitive loss ($\epsilon = 0.01$) and a Gaussian radial basis function (RBF) kernel with $\gamma = 0.001$. We used the function `svm` in package `e1071` for `R`, and set the `cost` parameter to 0.125. We compare the SVM with a penalized cubic spline as implemented in the package `SemiPar`. The optimal penalty term obtained with these data was `spar=63.4` and we kept it fixed throughout the rest of our analysis. Panels (a) and (c) in Fig. 16.1 contain 200 bootstrapped fits for each estimator.

We then added 6 outliers (around 2.5 % of atypical observations) and recomputed both estimators on 200 bootstrap samples. The tuning parameters were kept fixed at the same values used with the “clean” data. We display both sets of bootstrapped estimators (with and without outliers) for each estimator in Figs. 16.1(b) and 16.1(d). Darker lines correspond to bootstrapped estimators computed with outliers present in the data. Note that the bootstrap estimator of the distribution of the SVM estimates barely changes when outliers are present in the data. Penalized regression splines fits, however, are much more sensitive. Of course, these plots only represent one, but a typical, realization of the bootstrap estimate of the distribution of these regression methods with and without a small proportion of outliers. However, they serve to illustrate the different degrees of sensitivity of both methods to the presence of outliers.

16.5 Conclusions

Hable and Christmann (2011) showed that support vector machines based on the combination of (i) a continuous loss function, which is Lipschitz continuous and convex with respect to its third argument, and (ii) a bounded continuous kernel are qualitatively robust for any regularizing parameter $\lambda > 0$.

The main goal of this paper was to show that even the sequence of *bootstrap approximations* of SVMs based on such a combination of a loss function and a



(a) Bootstrapped SVMs on the original data

(b) Bootstrapped SVMs on both data sets

(c) Bootstrapped penalized cubic splines on the original data

(d) Bootstrapped penalized cubic splines on both data sets

Fig. 16.1 Illustration of the stability of bootstrapped support vector machines when the data contain a small proportion of outliers. *Lighter lines* correspond to bootstrap replicates on the clean data set, while *darker lines* are those obtained on the contaminated data set

kernel are qualitatively robust for $\lambda > 0$. Because the finite sample distribution of SVMs is unknown, approximations of the finite sample distribution are needed to construct approximate statistical tests or approximate confidence regions based on SVMs in this purely nonparametric setup. We hope that our results will stimulate further research on bootstrap approximations of SVMs or of related kernel based methods from machine learning theory.

Of course, qualitative robustness is just one notion of statistical robustness or of stability. There exist too many publications on robustness or stability and their relationship to learnability to mention all of them. Here we can only list a small selection of relevant papers on these topics.

The boundedness of the sensitivity curve of SVMs for classification was essentially already established by Bousquet and Elisseeff (2002). For results on influence functions and related quantities we refer to Christmann and Steinwart (2004) and Christmann and Steinwart (2007). If attention is restricted to SVMs based on a Lipschitz continuous convex loss function for regression and a continuous bounded kernel, Christmann and Van Messem (2008) showed that many SVMs even have a bounded Bouligand influence function in the regression context.

There is quite some interest in establishing relationships between stability, learnability, predictivity, and localizability for broad classes of regularized empirical risk minimization methods. We would like to mention Bousquet and Elisseeff (2002), Poggio et al. (2004), Mukherjee et al. (2006), and Elisseeff et al. (2005). Although different notions of stability are used in these papers, their notions of stability have a meaning similar to robustness in the sense of robust statistics. Several of these notions of stability only measure the impact of just *one* data point such that the connection to Tukey's sensitivity curve is obvious. Caponnetto and Rakhlin (2006) consider stability properties of empirical risk minimization over Donsker classes. For an interesting relationship between consistency and localizability we refer to Zakai and Ritov (2009).

Acknowledgements We thank PD Dr. Robert Hable and Dr. Thoralf Mildenberger for helpful discussions.

References

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, *2*, 499–526.

Caponnetto, A., & Rakhlin, A. (2006). Stability properties of empirical risk minimization over Donsker classes. *Journal of Machine Learning Research*, *7*, 2565–2583.

Christmann, A., Salibían-Barrera, M., & Van Aelst, S. (2011). On the stability of bootstrap estimators. arXiv:1111.1876.

Christmann, A., & Steinwart, I. (2004). On robust properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, *5*, 1007–1034.

Christmann, A., & Steinwart, I. (2007). Consistency and robustness of kernel based regression. *Bernoulli*, *13*, 799–819.

Christmann, A., & Van Messem, A. (2008). Bouligand derivatives and robustness of support vector machines for regression. *Journal of Machine Learning Research*, *9*, 915–936.

Christmann, A., Van Messem, A., & Steinwart, I. (2009). On consistency and robustness properties of support vector machines for heavy-tailed distributions. *Statistics and Its Interface*, *2*, 311–327.

Cucker, F., & Zhou, D. X. (2007). *Learning theory: an approximation theory viewpoint*. Cambridge: Cambridge University Press.

Cuevas, A. (1988). Qualitative robustness in abstract inference. *Journal of Statistical Planning and Inference*, *18*, 277–289.

Cuevas, A., & Romo, R. (1993). On robustness properties of bootstrap approximations. *Journal of Statistical Planning and Inference*, *2*, 181–191.

De Vito, E., Rosasco, L., & Toigo, A. (2012). *Learning sets with separating kernels*. arXiv:1204.3573.

Dudley, R. W. (1999). *Uniform central limit theorems*. Cambridge: Cambridge University Press.

Dudley, R. W. (2002). *Real analysis and probability*. Cambridge: Cambridge University Press.

Dudley, R. M., Giné, E., & Zinn, J. (1991). Uniform and universal Glivenko–Cantelli classes. *Journal of Theoretical Probability*, *4*, 485–510.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, *7*, 1–26.

Efron, B. (1982). *CBMS monograph: Vol. 38. The jackknife, the bootstrap, and other resampling plans*. Philadelphia: SIAM.

Elisseeff, A., Evgeniou, T., & Pontil, T. (2005). Stability of randomized learning algorithms. *Journal of Machine Learning Research*, *6*, 55–79.

Hable, R. (2012). Asymptotic normality of support vector machines variants and other regularized kernel methods. *Journal of Multivariate Analysis*, *106*, 92–117.

Hable, R., & Christmann, A. (2011). Qualitative robustness of support vector machines. *Journal of Multivariate Analysis*, *102*, 993–1007.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, *42*, 1887–1896.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the 5th Berkeley symposium* (Vol. 1, pp. 221–233).

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Ledoux, M., & Talagrand, M. (1991). *Probability in Banach spaces*. Berlin: Springer.

Mukherjee, S., Niyogi, P., Poggio, T., & Rifkin, R. (2006). Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, *25*, 161–193.

Poggio, T., Rifkin, R., Mukherjee, S., & Niyogi, P. (2004). General conditions for predictivity in learning theory. *Nature*, *428*, 419–422.

R Core Team (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org/.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge: Cambridge University Press.

Schölkopf, B., & Smola, A. J. (2002). *Learning with Kernels. Support vector machines, regularization, optimization, and beyond*. Cambridge: MIT Press.

Scovel, C., Hush, D., Steinwart, I., & Theiler, J. (2010). Radial kernels and their reproducing kernel Hilbert spaces. *Journal of Complexity*, *26*, 641–660.

Smale, S., & Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, *26*, 153–172.

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. New York: Springer.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. Singapore: World Scientific.

Talagrand, M. (1987). The Glivenko–Cantelli problem. *Annals of Probability*, *15*, 837–870.

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes: with applications to statistics*. New York: Springer.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.

Wendland, H. (2005). *Scattered data approximation*. Cambridge: Cambridge University Press.

Zakai, A., & Ritov, Y. (2009). Consistency and localizability. *Journal of Machine Learning Research*, *10*, 827–856.

Chapter 17
Some Machine Learning Approaches to the Analysis of Temporal Data

Katharina Morik

17.1 Introduction

The analysis of temporal data is an important issue in current research, because most real-world data either explicitly or implicitly contain some information about time. Time-related data include time series (i.e., equidistant measurements of one process), episodes made of events from one or several processes, and time intervals which are related (e.g., an interval overlaps, precedes, or covers another interval).

Statistical time series analysis has developed two big classes of representations, namely those in the time domain and those in the frequency domain. Analysis in the time domain is based on the correlation between the current and previous observations, while the frequency domain tries to decompose the time series into cyclic components at different frequencies. Time series are most often analyzed with respect to a prediction task, but also trend and cycle recognition belong to the statistical standard (see, for an overview, Schlittgen and Streitberg 2001).

In contrast to statistical time series analysis, data mining analyzes very large collections of time series. Indexing of time series according to similarity is needed for handling such collections (Keogh and Pazzani 2000) and has led to numerous representation methods for time series (e.g., Ding et al. 2008). Clustering of time series is a related topic (cf., e.g., Oates et al. 2001) as is time series classification (cf., e.g., Geurts 2001).

Event sequences are investigated in order to predict events or to discover patterns expressed by correlations of events. Mannila et al. (1997) define episodes:

> "An episode is a collection of events that occur relatively close to each other in a given partial order. We consider the problem of discovering frequently occurring episodes in a sequence. Once such episodes are known, one can produce rules for describing or predicting the behavior of the sequence."

K. Morik (✉)
Computer Science VIII, TU Dortmund University, 44221 Dortmund, Germany
e-mail: katharina.morik@tu-dortmund.de

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_17, © Springer-Verlag Berlin Heidelberg 2013

The approach of Höppner (2002) abstracts time series to time intervals and uses the time relations of Allen (1984) in order to learn episodes. The underlying algorithm is APRIORI by Agrawal et al. (1993) for learning frequent sets. The resulting episodes are written as association rules. Also inductive logic programming can be applied. Episodes are then written as logic programs, which express direct precedence by chaining unified variables and other time relations by additional predicates (Klingspor and Morik 1999). This requires some sort of abstraction or discretization. The abstraction of time series into sequences of events or time intervals approximates the time series piecewise by functions (so do, e.g., Keogh and Pazzani 1998, cf. Sect. 17.7).

There is a multitude of learning tasks related to temporal phenomena and, correspondingly, there are many possible representations for temporal data. Learning and representations are closely related: the *No Free Lunch Theorem* of Wolpert and Macready (1997) implies that finding an adequately biased representation can make a hard learning problem easy (and, vice versa, that finding this representation itself is hard). Morik (2000) focuses on the problem of selecting appropriate representations for time phenomena. In general, we have the following options:

- *Snapshot:* We ignore the time information and reduce the data to the most current state. This state can be written as one or several events. It may well happen that such a snapshot already suffices for learning (as in Sect. 17.2.2, for instance).
- *Value series:* Multi- or univariate series of numerical attributes, possibly stored in a number of records each showing the measurements for a time window of fixed length, which is shifted over the series (cf. Sect. 17.2.1). This representation is the starting point for most further processing.
- *Events with time intervals:* We aggregate time points to time intervals where attribute values are similar enough (piece-wise segmentation). For nominal attributes, it is straight forward to construct time intervals from the start and end time of each attribute value. For numerical attributes, feature extraction has to be performed first. In addition, we might want to represent relations between the intervals. Learning algorithms which make good use of time information (episode learning) can then be applied. Variants of this representation are illustrated in Sects. 17.3 and 17.6.
- *Feature extraction:* Time aspects are encoded as regular attributes of the examples such that any learning algorithm can be applied. Simple encodings are seasons simply stated by flags and attributes like *action_before* that summarize the events preceding a target event (cf. Sects. 17.2.1 and 17.5). More sophisticated methods apply transformations or functions (cf. Sect. 17.4).

In the following, case studies from time-related data analysis are presented. The chapter starts with using the Support Vector Machines as the learning method working on different representations. This is introduced in Sect. 17.2.1 and illustrated by an example using the snapshot representation in Sect. 17.2.2. Section 17.3 shows the importance of finding the appropriate representation of the temporal data: regardless which learning method was chosen, a smart transformation of the data allowed to successfully predict very rare events. The importance of an appropriate representation is stressed even further in Sect. 17.4, where those features are automatically extracted from time series that are well suited for classifying the set of time series at hand. The following Sections describe work that uses logic-based or relational learning methods. Section 17.5 handles the classification of phases and the use of logic-based learning for the analysis of concept shift. The other application area of a logic-based representation is robotics (Sect. 17.6). The relations of time intervals are applied to streaming sensor data, abstracting them to high level descriptions. For streaming data, new algorithms for a certain model have to be developed. In Sect. 17.7 an algorithm for streaming data is presented that detects patterns of flexible length.

Although being quite complementary, the overview of our work on time phenomena is devoted to Ursula Gather and her work on robust time series analysis, because collaboration has been always stimulating.

17.2 Support Vector Machines

Support Vector Machines (SVMs) are a well known prediction method (Vapnik 1998). They are handled in this book in the contribution by Christmann, Salibian-Barrera and van Aelst, Chap. 16. Before we show how this method can be used for handling time phenomena and illustrate it by a medical case study, we shortly introduce our notation. In its basic form, an SVM finds the hyperplane that separates the training data according to their label $y \in \{+1, -1\}$ with maximum margin. The learned model classifying vectors of observations $\mathbf{x}$ is written $f(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$. In order to maximize the distance between the examples that are closest to the separating hyperplane, called support vectors, the norm of the normal vector $\|\mathbf{w}\|$ needs to be minimized. The *soft margin* SVM weights the penalty term ξ for misclassified examples by a parameter C. For dealing with very unbalanced numbers of positive and negative examples, we introduce cost factors C_+ and C_- to be able to adjust the cost of false positives vs. false negatives. Finding this hyperplane can be translated into the following optimization problem:

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C_+ \sum_{i:y_i=1} \xi_i + C_- \sum_{j:y_j=-1} \xi_j \\ &\text{subject to} \quad \forall k : y_k[\mathbf{w} \cdot \mathbf{x}_k + b] \geq 1 - \xi_k. \end{aligned}$$

$\mathbf{x}_i$ is the feature vector of example i. y_i is the class label $+1$ or -1. $\mathbf{w}$ is the normal vector to the hyperplane. Since the optimization problem from above is difficult to handle numerically, it is transformed into its dual.

$$\begin{aligned} &\text{Minimize} \quad -\sum_{k=1}^{n} \alpha_k + \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \\ &\text{subject to} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned}$$

$$\forall i \text{ with } y_i = 1 : 0 \leq \alpha_i \leq C_+$$
$$\forall j \text{ with } y_j = -1 : 0 \leq \alpha_j \leq C_- .$$

The kernel function here is the linear kernel, i.e., the dot product $\mathbf{x}_i \cdot \mathbf{x}_j$. Other kernel functions, such as, e.g., the radial basis function, can be used as well. We solve this optimization problem using SVM$^{\text{light}}$ of Joachims (1999), extended to handle asymmetric cost-factors. It can efficiently handle problems with many thousand support vectors, converges fast, and has minimal memory requirements.

17.2.1 Support Vector Models of Time Series

SVMs have been applied to time series prediction (e.g., Müller et al. 1997) based on a value series representation, i.e., creating d-dimensional examples by moving a window of length d over the time series. However, the theory of structural risk minimization, on which the SVM is based, is only formulated for independent, identically distributed data. Clearly, the independence assumption is violated for time series data. Using the value series representation with a linear model leads to the class of autoregressive (AR) models (Schlittgen and Streitberg 2001). Obviously, AR models can be learned by an SVM with linear kernel, so it does not surprise that the SVM does not perform very different on data generated from an AR model than other methods for AR model estimation. For time series analysis in the frequency domain, the Fourier transformation can be used to transform the examples for the SVM. There also exist kernel functions, which perform this transformation, e.g., the Fourier kernels as introduced by Vapnik (1998) or the time-frequency kernel of Davy et al. (2002).

The main advantage of SVMs is that the generalization error does not depend on the dimensionality of the data but only on the margin of the separating hyperplane, which makes them especially well suited for high-dimensional data. While this reasoning is not strictly valid—the margin depends on the geometry of the data and hence also on the dimension—empirical evidence shows that this property of SVMs does hold in practice. Hence, adding attributes that express temporal phenomena is encouraged. Seasonal attributes are a good example. The case of cyclical components is more complicated, because they cannot be as easily identified and filtered from the data as are simple trends. Often the most difficult problem in practice is that lots of statistical procedures for modeling periodic functions cannot be applied, because what looks like a periodic component actually is not one. Therefore, in Rüping (1999), 20 additional binary attributes were used to mark the presence of holidays, special sale promotions, and other significant events in that particular week.

The moving window representation bases on the assumption that the temporal dependence structure of the time series can be sufficiently captured in a short finite window of observations. This allows the examples generated from each window to be treated as if they were generated independently. This assumption fails, if the process that generates the time series changes over time. This scenario is called concept drift. Usually, a concept drift is treated by using only a certain number of the newest examples, where the actual number of examples used is chosen heuristically. For SVMs, Klinkenberg and Joachims (2000) propose an approach where this number is chosen based on efficient performance estimators for SVMs.

17.2.2 Intensive Care—A Case Study

In modern intensive care, several hundreds of measurements of a patient are recorded at the bed-side each minute or even second. This results in masses of noisy, high dimensional, sparse time series of numerical data (for an overview, see Fried et al. 2000). A collaboration between Ursula Gather from statistics and Michael Imhoff from Medicine dealt with data collected at the 16-bed intensive care unit (ICU) of the "Chirurgische Kliniken der Städtischen Kliniken Dortmund". The particular data set that we used contains the data of 147 patients between January 1997 and October 1998. Measurements are taken every minute, amounting to 679,817 observations for which data from a Swan-Ganz catheter is available. Real valued parameters are either scaled so that all measurements lie in the interval $[0, 1]$, or they are normalized by empirical mean and variance. There are 118 attributes forming 9 groups and here we use just 8 vital signs of the hemodynamic system. In addition, the dose rates of 6 drugs were also recorded.

The task of monitoring can best be understood as time-critical decision support. The final goal is to enhance the quality of clinical practice. This means that imitating the actual interventions, i.e., the doctor's behavior, is not the goal. It is instead to supply physicians with the best recommendation under all circumstances (Morris 1998). Finding alarm functions which minimize the number of false positives while keeping the number of false negatives small has been solved by Sieben and Gather (2007). Alarm functions need to robustly extract the patient's state from time series as is investigated in the contribution by Borowski, Fried and Imhoff in Chap. 12. Determining the appropriate action for a patient's state is a further task of decision support and is described in the following.

Learning State-Action Rules We have split the acquisition of state-action rules into two learning tasks. The first task was to learn a function on whether a certain drug should be given to a particular patient, or not. Positive examples are the data of all those minutes, in which the drug was given. All others are negative examples. The second task is to learn functions on whether the dose of the drug should be increased, decreased, or kept constant. This task was transformed into three two-class learning problems. All experiments towards finding an appropriate representation (feature selection) and on optimizing the parameters of the SVM were done on the training set using 10-fold cross validation. Comparing the use of different kernel functions led to using linear SVMs for all drugs (cf. Morik et al. 1999, 2002).

The learned function for a particular drug is applied to a patient's current state in terms of $\text{svm\_calc}(x) = \sum_{i=1}^{p} \beta_i x_i$ which is then turned into a binary decision.

In the first task, for example, the patient pat45 in minute 25 with arterial pressures 174 (systolic), 86 (diastolic), 121 (mean) and pulmonal pressures 26 (systolic), 13 (diastolic), 15 (mean), heart rate 79 and central venous pressure 8 with svm_calc(pat45) $= 7.638$ would receive the recommendation of not taking the drug, because $\mathrm{sign}(\mathrm{svm\_calc}(x) - \beta_0)$ is negative for $\beta_0 = 4.368$.

$$f(\mathbf{x}) = \left[\begin{pmatrix} 0.014 \\ 0.019 \\ -0.001 \\ -0.015 \\ -0.016 \\ 0.026 \\ 0.134 \\ -0.177 \end{pmatrix} \begin{pmatrix} \mathrm{artsys} = 174.00 \\ \mathrm{artdia} = 86.00 \\ \mathrm{artmn} = 121.00 \\ \mathrm{cvp} = 8.00 \\ \mathrm{hr} = 79.00 \\ \mathrm{papsys} = 26.00 \\ \mathrm{papdia} = 13.00 \\ \mathrm{papmn} = 15.00 \end{pmatrix} - 4.368 \right].$$

To get an impression about how good the predictions are, we conducted an experiment with an experienced ICU physician. On a subset of 40 test examples we asked the expert to do the same task as the SVM for Dobutamin, given the same information about the state of the patient. In a blind test he predicted the same direction of dosage change as actually performed in 32 out of the 40 cases. On the same examples, the SVM predicted the same direction of dosage change as actually performed in 34 cases, resulting in an essentially equivalent accuracy. The SVM and the expert agreed in all the 32 cases. Hence, the learned function follows the observations more closely than the expert who may have reasons to deviate from them.

The second task is more detailed. Given the state of the patient, should the dosage of a drug be increased, decreased or kept constant? Knowing such a function is also a step towards deciding when to substitute one drug with another. To make sure that we generate examples only when a doctor was closely monitoring the patient, we considered only those minutes where some drug was changed, resulting in 1319 training and 473 test examples. For each drug, we trained two binary SVMs. One is trained on the problem "increase dosage" vs. "lower or keep dosage equal", the other one is trained on the problem "lower dosage" vs. "increase or keep dosage equal".

For a subsample of 95 examples from the test set, we again asked the medical expert to perform the same task as the SVM. The results for Dobutamin and Adrenalin are given in Table 17.1. The performance of the SVM on this subsample is followed by the performance of the human expert (in brackets). Both are well aligned. Again, the learned functions of the SVM are comparable in terms of accuracy with a human expert. This also holds for the other drugs.

Summarizing the case study in intensive care, learning recommendations of treatments by a linear SVM was done offline using a *snapshot* of patient data. The complex learning task was divided into one learning task for each drug and three learning tasks for the dose of each drug, i.e., 24 learning tasks were performed. The result is a model that is applied every minute: reading in the current state, applying all learned decision functions, in parallel, thus delivering the recommendations. Hence, the application of the model follows already the real-time processing of data streams.

Table 17.1 Confusion matrices for predicting time and direction of Dobutamin and Adrenalin interventions in comparison to human performance (human performance in *brackets*)

Dobutamin	Actual intervention			Adrenalin	Actual intervention		
	up	equal	down		up	equal	down
predicted up	**10** (**9**)	12 (8)	0 (1)	predicted up	**4** (**2**)	3 (1)	0 (0)
predicted equal	7 (9)	**35** (**31**)	9 (9)	predicted equal	4 (6)	**65** (**66**)	2 (2)
predicted down	2 (1)	7 (15)	**13** (**12**)	predicted down	1 (1)	8 (9)	**8 (8)**

17.3 Temporal Databases—An Insurance Case Study

Time-stamped data are observations with an attached time-stamp. They occur frequently in real-world databases, where dates, hours, minutes are common attributes. Often, several rows in a database table describe the same object, each row for one of the object's states. Usually, two attributes for the beginning and the end point in time are part of the row (vector) describing an object's state. Whenever an attribute's value has changed, a new row is added.

- The snapshot approach would just extract the most current row for an object.
- The time interval approach would use the "begin" attribute and the "end" attribute and indicate the other attributes from a row as an event.

For the frequency count of an attribute, we simply count how often its value changes. This representation is similar to the term frequency as used in information retrieval for texts. Analogously to the representation there, we exclude the frequencies of those changes that are common to all objects.

In a study for the Swiss Life insurance company, we investigated the time-stamped data of customers and their contracts. In the course of enhanced customer relationship management, the Swiss Life insurance company investigated opportunities for direct marketing (see Kietz et al. 2000). A more difficult task was to predict surrender in terms of a customer buying back his life insurance. We worked on knowledge discovery for the classification into early termination or continuation of policies. The task was clearly one of local pattern learning: only 7.7 % of the contracts end before their end date. Hence, the event to be predicted is rare. Internal studies at the insurance company found that for some attributes the likelihood of surrender differed significantly from the overall likelihood. In each contract, there are several attributes indicating surrender or continuation. We also found that within the group of terminated contracts, there were those which do not share attributes. This reminds us of the characteristics of text classification as investigated by Joachims (2002). There, features from information retrieval express a pair of orthogonal frequencies: term frequency in one document and inverse document frequency for a term (TFIDF) introduced by Salton and Buckley (1988). This leads to a third possibility for representing temporal information.

- If we transform the raw data into a frequency representation, we possibly condense the data space in an appropriate way.

For the frequency count of an attribute, we simply count how often its value changes. In the Swiss Life application, term frequency tf describes how often a particular attribute a_i of the contract or one of its components c_j has been changed:

$$\mathrm{tf}(a_i, c_j) = \left\| \left\{ x \in \text{time points} \mid a_i \text{ of } c_j \text{ changed} \right\} \right\|.$$

The document frequency here corresponds to the number of contracts in which a_i has been changed. The set of all contracts is written C. The document frequency df is just the number of contracts with a term frequency greater than 0:

$$\mathrm{df}(a_i) = \left\| \left\{ c_j \in C \mid a_i \text{ of } c_j \text{ changed} \right\} \right\|.$$

Hence, the adaptation of the TFIDF text feature to contract data becomes for each contract c_j:

$$\mathrm{tfidf}(a_i) = \mathrm{tf}(a_i, c_j) \log \frac{\|C\|}{\mathrm{df}(a_i)}.$$

A given anonymous database consists of 12 tables with 15 relations between them. The tables contain information about 217,586 policies and 163,745 customers. If all records referring to the same policy and component (but at a different status at different times) are counted as one, there are 533,175 components described in the database. We selected 14 attributes from the original database. 13 of them were transformed as described above. One of them is the reason for a change of a contract. There are 121 different reasons. We transformed these attribute values into binary attributes. Thus we obtained $13 + 121 = 134$ features describing changes of a contract. To calculate the TFIDF values for these binary features we considered the history of each contract. For the 121 newly created features we counted how often they occurred within the mutations. With this representation we could predict whether a contract is bought back by a customer. We compared the learning results on this generated representation to those on the selected original data for different learning algorithms. We used 10-fold cross validation on a sample of 10,000 examples. In order to balance precision and recall, we used the F-measure

$$F_\beta = \frac{(\beta^2 + 1)\,\mathrm{Prec}(h)\,\mathrm{Rec}(h)}{\beta^2\,\mathrm{Prec}(h) + \mathrm{Rec}(h)},$$

where β indicates the relative weight between precision Prec and recall Rec, and h the learned decision function. We have set $\beta = 1$, weighting precision and recall equally. For all algorithms, the frequency features are better suited than the original attributes. For the SVM, the performance changed from a low F-measure of 16.06 % using the original representation to 97.95 % using the TFIDF representation.

Summarizing the case study of time-stamped insurance data, we underline that likelihood estimation was not enough, but we needed to learn a model, which can be readily used to act, i.e. send an insurance agent to the customer. The skewed distribution made this very difficult. Only when we extracted features, similar to term frequency and inverse document frequency, thus counting the frequency of contract changes, the classification became very precise. Extracting this *TFIDF feature* is very often a good opportunity for learning from time-stamped data. Moreover, Joachims' theory of text classification using the SVM computes a tight upper bound of the error using only one SVM run (Joachims 2002). Hence, the error could be estimated by the tight bound.

17.4 Classifying Time Series—A Music Mining Case Study

For classification, machine learning encounters a challenge of scalability, when confronted with music data. Music databases store millions of records. Given a sampling rate of 44,100 Hz, a three minute music record has the length of about 8×10^6 values. Moreover, current approaches to time series indexing and similarity measures rely on a more or less fixed time scale (e.g., Keogh and Pazzani 1998). Music plays, however, differ considerably in length. More general, time series similarity is often determined with respect to some (flexible and generalized) shape of curves as stated by Yi et al. (1998). However, the shape of the audio curve does not express the crucial aspect for classifying genres or preferences. The i-th value of a favorite song has no correspondence to the i-th value of another favorite, even if relaxed to the $(i \pm n)$-th value. The decisive features for classification have to be extracted from the original data. Some approaches extract features from music that is represented in form of Midi data, i.e., a transcription according to the 12 tone system (cf. Weihs and Ligges 2005). This allows to include background knowledge from music theory. The audio data are given, however, in the form of—possibly compressed—waves records. Hence, feature extraction from audio data has been investigated, e.g., by Tzanetakis (2002). Several specialized extraction methods have shown their performance on some task and data set. It is now hard to find the appropriate feature set for a new task and data set. In particular, different classification tasks ask for different feature sets. It is not very likely that a feature set delivering excellent performance on the separation of classical and popular music works well for the separation of techno and hip hop music, too. Classifying music according to user preferences even aggravates the problem. Hence, for every learning task, we have to search in the space of possible feature extractions. This search can only be automated if the search space is well structured.

Audio data are time series where the y-axis is the current amplitude corresponding to a loudspeaker's membrane and the x-axis corresponds to the time. They are univariate, finite, and equidistant. We may generalize the type of series which we want to investigate to *value series*. Each element x_i of the series consists of two components. The first is the *index component*, which indicates a position on a straight line (e.g., time). The second component is an m-dimensional vector of values which is an element of the *value space*. We can now structure the set of elementary operators which allow to compose all possible feature extractions.

- *Basis transformations* map the data from the given vector space into another space, e.g., frequency space, function space, phase space. The most popular transformation is the Fourier analysis.

- *Filters* transform elements of a given series to another location within the same space. The moving average or exponential smoothing are examples of filters.
- *Mark-up of intervals* corresponds to the mark-up of text fragments in that it annotates segments within a value series.
- *Generalized windowing* is required by many methods for feature extraction. We separate the windowing from the functions applicable to values within the windows.
- *Functions* calculate a single value for a value series. Typical examples are average, variance, and standard deviation.

Since the group of mark-up operators is newly introduced, a definition is given here.

Definition 17.1 (Mark-up) A mark-up $M : S \rightarrow C$ assigns a characteristic C to a segment S.

Definition 17.2 (Interval) An interval $I : S \rightarrow C$ is a mark-up within one dimension. The segment $S = (d, s, e)$ is given by the dimension d, the starting point s, and the end point e. The characteristic $E = (t, \varrho)$ indicates a type t and a density ϱ.

Operators finding intervals in the value dimension of a value series can be combined with the mark-up of intervals in the time (i.e., indexing) dimension. For instance, whenever an interval change in the value dimension has been found, the current interval in the index dimension is closed and a new one is started.

Many known operators on times series involve windowing. Separating the notion of windows over the index dimension from the functions applied to the values within the window segment, allows to construct many operators of the kind.

Definition 17.3 (Windowing) Given the series $\mathbf{x}$ with $i \in \{1, \ldots, n\}$, a transformation is called windowing, if it shifts a window of width w using a step size of s and calculates in each window the function F: $y_j = F(\mathbf{x}_i)$ with $i \in \{j \cdot s, \ldots, j \cdot s + w\}$. All y_j together form again a series $\mathbf{y}_j$ with $j \in \{1, \ldots, (n - w)/s + 1\}$.

Definition 17.4 (General Windowing) A windowing which performs an arbitrary number of transformations in addition to the function F is called a general windowing.

The function F summarizes values within a window and thus prevents general windowing from enlarging the data set too much. Since the size of audio data is already rather large, it is necessary to consider carefully the number of data points which is handled more than once. The *overlap* of a general windowing with step size s and width w is defined as $g = w/s$. Only for windowings with overlap $g = 1$ the function can be omitted. Such a windowing only performs transformations for each window and is called *piecewise filtering*. Combining general windowing with the mark-up of intervals allows to consider each interval being a window. This results

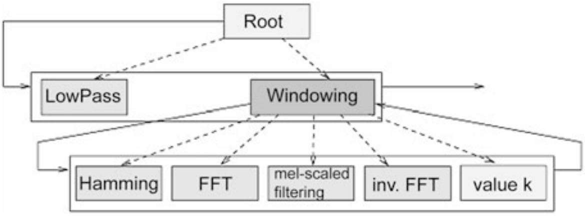


Fig. 17.1 Constructing the cepstral method from elementary extraction operators

in an adaptive window width w and no overlap. Of course, this speeds up processing considerably.

The elementary operators can be combined so that methods of feature extraction are expressed. The *mel-frequency cepstral coefficients* can be constructed as a general windowing, where the frequency spectrum of the window is calculated, its logarithm is determined, a psychoacoustic filtering is performed, and the inverse Fourier transformation is applied to the result. Figure 17.1 shows how the operators for feature extraction are put together to form the cepstral coefficients. From these coefficients, additional features can be extracted. It is seen how easily new, similar methods can be generated, e.g., by replacing the frequency spectrum and its logarithm by the gradient of a regression line.

This general framework integrates a large variety of standard functions and offers the opportunity to construct new features that are tailored for an application. Mierswa and Morik (2005) have shown that evolutionary learning automatically constructs features for diverse tasks of music mining. A separate set of features is constructed for each particular application, ranging from the preferences of a particular user, to classes of mood, genre, instrumentation, or for whatever concept. The method can be applied to learn features for any classification task, where a set of positive and negative examples can be presented. This learning capability is a basis of further services to users of large music collections developed by Wurst et al. (2006).

Using the automatic feature construction, genre classification and the classification of user preferences could successfully be learned. For the classification of genres, three data sets have been built: for classic versus pop, 100 pieces of music for each class were available, for techno versus pop, 80 songs for each class from a large variety of artists were available, and for hiphop versus pop, 120 songs for each class from few records were available. The classification tasks are of increasing difficulty. Using mySVM (Rüping 2000) with a linear kernel, the performance was determined by a 10-fold cross validation and is shown in Table 17.2. Concerning classic vs. pop, 93 % accuracy, and concerning hiphop vs. pop, 66 % accuracy have been published by Tzanetakis (2002). 41 features have been constructed for all genre classification tasks. For the distinction between classic and pop, 21 features have been selected for mySVM by the evolutionary approach. For the separation of techno and pop, 18 features were selected for mySVM, the most frequently selected ones being the filtering of those positions in the index dimension where the curve crosses the zero line. For the classification into hiphop and pop, 22 features were selected with the mere volume being the most frequently selected feature.

Table 17.2 Classification of genres with a linear SVM using the task specific feature sets

	Classic/pop	Techno/pop	Hiphop/pop
Accuracy	100 %	93.12 %	82.50 %
Precision	100 %	94.80 %	85.27 %
Recall	100 %	93.22 %	79.41 %
Error	0 %	6.88 %	17.50 %

Table 17.3 Classification according to user preferences

	User_1	User_2	User_3	User_4
Accuracy	95.19 %	92.14 %	90.56 %	84.55 %
Precision	92.70 %	98.33 %	90.83 %	85.87 %
Recall	99.00 %	84.67 %	93.00 %	83.74 %
Error	4.81 %	7.86 %	9.44 %	15.45 %

Another task requiring feature extraction is the recommendation of songs to particular users. The classification of user preferences beyond genres is a challenging task, where for each user the feature set has to be learned. Four users brought 50 to 80 pieces of their favorite music ranging through diverse genres. They also selected the same number of negative examples. Using a 10-fold cross validation, mySVM was applied to the constructed and selected features, one feature set per learning task (see Table 17.3). The excellent learning result for a set of positive instances which are all from a certain style of music corresponds to our expectation (user 1). The expectation that learning performance would decrease if positive and negative examples are taken from the same genre is not supported (user 2). Surprisingly well is the learning result for a broad variety of genres among the favorites (user 3). Sampling from only a few records made the learning task more difficult as is shown by the results for user 4.

Applying the learned decision function to a database of records allowed the users to assess the recommendations. They were found very reasonable. No particularly disliked music was recommended, but unknown plays and those, which could have been selected as the top 50.

In summary, the challenge of music mining lies in the very large collection of values series, where the similarity between different instances needs to be determined by new variables that are based on *feature extraction*. The same functionality can as well be used for other *value series*. Video classification or clustering are other examples, but it is not restricted to multimedia applications. For example, Ritthoff et al. (2002) solved a regression problem for time series with SVMs, where certain coefficients of chemical components have been predicted from chromatography time series. Even without evolutionary optimization, the structured set of operators eases time series analysis. Within the tool *RapidMiner* developed by Mierswa et al. (2006), the series extension provides many series transformation and extraction operators, which make regression or classification applicable.

17.5 Logic Rules and Concept Shift—A Business Cycles Case Study

This case study illustrates the use of logic representations for temporal cycles. Here, we use the interval relations of Allen (1984). The application is an economic study of German business cycles. We were given quarterly data for 13 economic indicators concerning the German business cycle from 1955 to 1994, where each quarter had been classified as being a member of one of four phases, namely *up, upper turning point*, *down, lower turning point* (cf. Heilemann and Munch 2001). The ups and downs of business activities have been observed for a long time. It is, however, hard to capture the phenomenon by a clear definition. The National Bureau of Economic Research (NBER) defines business cycles as "recurrent sequences of altering phases of expansion and contraction in the levels of a large number of economic and financial time series". The learning task of *dating* is defined as classifying current measurements into the phases of a business cycle.

Before rule learning in logic programming can be applied, the originally real-valued time series of indicator values have to be transformed into discrete-valued temporal facts about these indicators. Using the information that is given by the examples' class (the business phase) improves discretization. In this case, we exploited the built-in discretization of C4.5 (Quinlan 1993). Inducing decision trees about the cycle phases, based on only one indicator Y derives split points for Y. The resulting trees were cut off at a given level and the decisions in this resulting tree were used as discretization thresholds. Decision trees of depth 2, i.e., using 4 discrete values, proved to build a suitable number of facts. We have used the discretization of the indicator values for the construction of time intervals. This representation abstracts from minor changes. We then learned rules in restricted predicate logic using the Rule Discovery Tool for inductive logic programming which is described in the book by Morik et al. (1993). Here, we do not present the learned logical rules; interested readers might see Morik and Rüping (2002). Instead, we want to show how to use sets of learned rules in order to discover concept shift.

Analyzing Concept Shift by Frequent Sets Concept shift is an important issue for serial data. For time series which express equidistant numerical measurements, rank tests are capable of identifying shifts (Fried and Gather 2007). In the business cycle application here, the concept shift is to detect from discrete facts. Since the four-phase model did not lead to convincing learning results, we used the two-phase model, which distinguishes only up and down and not the turning points. We analyzed the homogeneity of the business cycle data using the sets of learned rules. The learning results from different leave-one-cycle-out experiments using logic learning were inspected with respect to their correlation. Since inductive logic programming delivers all valid rules, we can infer from a rule not being learned that it is not valid. If the same logical rule is learned in all experiments, this means that the underlying principle did not change over time. If, however, rules co-occur only in the first cycles and not in the following ones, we hypothesize a concept drift in business cycles.

We used the correlation analysis of the APRIORI algorithm by Agrawal and Srikant (1994).

We want to know whether there are rules that are learned in all training sets, or, at least, whether there are rules that are more frequently learned than others. Enumerating all learned rules results in a vector for each training set (corresponding to a transaction in frequent set mining) where a learned rule is marked by 1 and those that are not learned are set to 0. The frequency of learned rules and their co-occurrence is identified. There is no rule which was learned in all training sets. Eight rules were learned from three training sets. No co-occurrence of learned rules was found. There is one rule, which was learned in four training sets, namely leaving out cycle 1, cycle 4, cycle 5, or cycle 6: $\mathrm{rld}(T, V), \mathrm{l}(T, V), \mathrm{low}(V) \rightarrow \mathrm{down}(T)$ stating that the same discrete value V of the real long term interest rate (rld) in phase T and of the number of wage and salary earners (l) in phase T being low (low) indicates that T is a downswing.

We now turn around the question and ask: which training sets share rules? For answering this question, a vector for each learned rule is formed where those training sets are marked by 1 which delivered the rule. The rule set analysis shows that cycles 1 to 4 (1958–1974) and cycles 3 to 6 (1967–1994) are more homogeneous than the overall data set. The first oil crisis happened at the end of cycle 4 (November 1973–March 1974). This explains the first finding well. However, the oil crisis cannot explain why cycles 3 to 6 share so many rules. We assume that the actual underlying rules of business cycles may have changed over time. The concept drift seems to start in cycle 3. The periods of cycles 1 and 2 (1958–1967) are characterized by the reconstruction after the world war. Investment in construction (ic) and in equipment (ie) is not indicative in this period, since it is rather high, anyway. A low number of earners (l) together with a medium range of the gross national product deflator (pyd) best characterizes the downswing in cycles 1 to 3—this rule has been found when leaving out cycles 4 or 5 or 6. Since the unemployment rate was low after the war, it is particularly expressive for dating a phase in that period. This explains the second finding of our rule set analysis. For more details of the learned rules, see Morik and Rüping (2002).

The case of inspecting business cycles shows how logic rules can be used to detect concept shifts. This is possible, because inductive logic programming delivers all valid rules. We made use of this in this case study. First, numerical values of economic features are discretized, then logical rules are learned. Here, we inspected cycles in sequences of *time intervals*. The logical rules are then analyzed with respect to *concept shift* using frequent set mining. This novel set-up complements methods for numerical data.

17.6 Logic-Based Learning and Streams—A Case Study in Robotics

Usually, if a robot has to navigate in office rooms, the commands for movements are given in terms of coordinates. However, offices are looking more or less the same wherever they are. Users should be allowed to give commands such as: *move through the doorway, turn left, move until the cupboard, and stop.* Learning all the way from the mobile robot's sensing to human communication and back to the robot's acting is composed of several learning tasks, each abstracting a level to the next higher one until concepts at the human level of abstraction are acquired. These concepts are operational in that they can be transformed easily into the robot's control. Rieger and Klingspor (1999) developed some improvements for real-time processing. The mobile robot PRIAMOS built by Kaiser and Dillmann (1999) was equipped with ultra-sonar sensors. How these distributed sensor streams are processed is described in Sect. 17.7.

A set of rules in a restricted predicate logic allows to infer the current state of the robot in rules that lead from the basic perception of the sensors to an operational concept and from the operational concept to actions of the robot. The rules are a logic program which can be executed. The distinction between features and concepts is usually established by their different representations. The features of an instance are represented by values in a vector, where the concept is represented by a set of rules. In restricted predicate logic, the distinction between features and concepts vanishes. A concept as well as a feature is represented by a predicate whose meaning is given by a set of rules: a concept is another's concept feature. This allows us to apply the same learning method at different levels of abstraction. Klingspor (1994) developed the algorithm GRDT based on RDT (Kietz and Wrobel 1992), which learns all valid rules of a form which is declaratively described by a grammar. For learning, paths of the robot in offices of the University of Karlsruhe served as observational data. The learned models were tested by controlling the robot in offices of the Technische Universität Dortmund. All commands were accomplished successfully.

The key to representing time is *unification* here. A rule is instantiated by sensor measurements and the current time stamp. In a rule, only the identity or non-identity of time points plays a role. When instantiated, unification puts the time points into order. The chaining temporal arguments in the logic predicates represent an order of events, T1 to T2, then T2 to T3, then T3 to T4, all summarized by the interval T1 to T4. An example showing the levels from an operational concept to the perception may illustrate this.[1] The rule describes the operational concept of moving through a doorway in a parallel manner.

Operational concept

standing (Trace, T1, T2, *in_front_of_door*, PDirect, small_side, PrevP) &
parallel_moving (Trace, T2, T3, MSpeed, PDirect, *through_door*,
right_and_left) &
standing (Trace, T3, T4, *in_front_of_door*, back, small_side, *through_door*)
→ move_through_door(Trace, T1, T4)

[1] We follow the Prolog convention and write constant terms in small letters, variables start with a capital letter. For easy reading, the perceptual features are written in italics.

The first premise states that the robot is standing and sensing with the sensors at its small side the perceptual feature *in_front_ of_door* in a time interval from T1 to T2. Next, from time point T2 to T3, the action *parallel_moving* is performed, measuring by the sensors at the right and the left side of the robot the perceptual feature *through_door*. Then, from T3 to T4, having passed the doorway is recognized by the perceptual feature *in_front_of_door* being sensed by the back sensors of the robot.

Some arguments of an operational concept refer to action-oriented perceptual features. These are defined by a hierarchy of rules. We may trace the feature *through_door* from its use as an argument in the concept *move_through_door* to its basic perceptual features.[2]

Action-oriented perceptual feature

```
sg_jump(Tr, right_side, T1, T2, Move) &
sg_jump(Tr, left_side, T3, T4, Move) &
parallel(Move) & succ(T1, T3) & Start ≤ T1 & T2 ≤ End
    → through_door (Tr, Start, End, Move).
```

The rule expresses that the door post is sensed by the sensors at the right side and a little later by the sensors at the left side, expressed by the successor relation between T1 and T3, the starting time points of the sensor feature *sg_jump*. The movement is in parallel, because the variable *Move* is unified for the right and the left side of the robot. This pattern is summarized by the predicate *through_door*. Similar rules are learned for other features from robot traces.

The perceptual feature *sg_jump* is defined in terms of basic perceptual features.

Perceptional feature

```
stable(Tr, Orien, S, T1, T2, Grad1) &
incr_peak(Tr, Orien, S, T2, T3, Grad2) &
stable(Tr, Orien, S, T3, T4, Grad3)
    → sg_jump(Tr, S, T1, T4, parallel)
```

This rule states that a (group of) sensors *S* in a particular orientation *Orien* measures a sequence of distance values that form in the time interval from T1 to T2 the gradient *Grad1*, in time interval from T2 to T3 the gradient *Grad2*, and in time interval from T3 to T4 the gradient *Grad3*. The end point of one time interval is the start point of the next. *stable* and *incr_peak* are basic features that need to be computed

[2] For technical reasons we had to introduce conversion rules that convert the predicate of a rule's conclusion into an argument of another predicate.

as streaming data in real-time (see Sect. 17.7). Unifying the variables with particular sensor facts derives the unified conclusion which is used as a premise in the action-oriented perceptual features. If its other premises could be derived as well, the action-oriented perceptional feature is turned into an argument of an action feature like *standing* or *parallel_moving*, thus, in the end contributing to the derivation of an operational concept.

Due to its relational representation of time intervals, operational concepts are flexible. The definition of *move_through_door* applies to doorways of varying breadth and depth. No precise distances need to be fixed by the definition. The same holds for time intervals. No fixed length has to be determined, but only the *relations between time intervals* and even these can freely be expressed. In summary, the case-study shows that handling time in the logical framework is very flexible. Moreover, the results of learning from very few training paths of a robot control the robot at other buildings.

17.7 Streaming Data Analysis—A Very Early Algorithm

Online time series analysis is an important method for monitoring tasks. It allows to analyze the time series without storing it completely. Hence, it is applicable to streaming data. Robust methods have recently been developed by Borowski et al. (2009), Nunkesser et al. (2009), and Lanius and Gather (2010) and are shown in the contribution by Borowski, Fried and Imhoff in Chap. 12. It was in the course of exploiting logic learning and reasoning in robotics, that we first came across streaming data that had to be handled online and under resource constraints regarding computing and memory. The task was to transform the sensor measurements, i.e., numerical streaming data, into symbolic descriptions. Since the technical term of *streaming data analysis* did not yet exist, we strengthened the notion of *incremental* becoming the definition of algorithms for streaming data (Morik and Wessel 1999):

> "If we want to use the symbolic descriptions for object recognition or planning during the robot's performance, then the transformation from sensor data to symbolic descriptions must be very fast. This not only means that fast algorithms are to be applied. It also means that the algorithm must be incremental, i.e. the algorithm generates the symbolic description on the basis of the current symbolic description and the current measurement. An algorithm is called *strongly incremental* if it does not store the input data so that it cannot correct its decisions on the basis of a longer input stream."

Moreover, we demanded an any-time algorithm: each current symbolic description must be informative to the robot. It is impossible to wait for a longer sequence of sensor data in order to get a better symbolic description. The robot needs information about its environment at any time.

We developed an algorithm that takes as input a set of symbolic labels, a sequence of distance measurements of sonar sensors, and a parameter that indicates the degree of tolerance concerning deviations. Its output is a sequence of labeled time intervals. For example, during a robots movement along a wall, a sensor measures an object first with increasing and then, shortly, with decreasing distance, because something juts out. The logic proposition *increasing(t100, 207, s0, 1,8,19)* is output at the ninth time point. It states that in the current mission *t100* the sensor *s0*, being oriented with respect to the global coordinates in an angle of 207 degrees, has measured increasing distances from the 1st time point until the 8th time point. The gradient of the increasing distances was about 19 degrees. At the ninth point of time, a new time interval is started, because the gradient is too different. It is labeled $decr\_peak(t100, 207, s0, 8, 9, -64)$.

For two measurements, we have the position of the robot Rx and Ry at two consecutive time points and the distance measurements D at the same time points. We are only interested in changes of distance measurements. Such a change is expressed by the gradient α between two measurements:

$$\alpha = \arctan \frac{D_t - D_{t-1}}{(Rx_t, Ry_t) - (Rx_{t-1}, Ry_{t-1})}.$$

For every two successive measurements the angle is calculated. For the first two measurements, the angle α is the *characteristic gradient* of the time interval. For every further measurement, its gradient is compared with the characteristic one. If the gradient is decided to be more or less the same, then the characteristic gradient is updated. The update forms incrementally the average of angles that are summarized by the time interval.

$$\text{char\_grad}_{\text{new}} = \frac{\text{char\_grad}_{\text{old}} \cdot (\#\text{grads} - 1) + \text{grad}}{\#\text{grads}}$$

calculates the new characteristic gradient $\text{char\_grad}_{\text{new}}$ from $\text{char\_grad}_{\text{old}}$, the current characteristic gradient, weighted by #grads, the number of gradients that have been subsumed so far, and the current gradient grad. In preparation of an application, the tolerance threshold which decides whether an angle s is similar enough to the characteristic gradient, is fitted to a training set once. From then on, the algorithm works in a streaming any-time manner. The signal to symbol processing could be used for real missions of the robot PRIAMOS given the previously learned tolerance parameter, successfully.

The little exercise for a mobile robot illustrates a very early algorithm of a class that now has attained much attention, namely, the streaming data algorithms. Sensor-based monitoring and prediction has become a hot topic in a large variety of applications, see for instance the contribution by Borowski, Fried and Imhoff in Chap. 12. Here, the robot sensor measurements were processed in parallel, one process for each sensor, before a learned model combined the distributed patterns. The learning of perceptional patterns had to store only the current model and the new measurement. The extracted features were then used by the logic rules. The strongly incremental segmentation and summary of on-line time series already instantiates the mining of streaming data from heterogeneous sources. The theoretical analysis of stream mining algorithms has also been put forward (for clustering, see Ackermann et al. 2010, for instance). Stream mining allows to handle a virtually infinite number of observations. Hence, it is an important method for analyzing *big data*.

17.8 Conclusions

Temporal data offer fascinating research topics. In this promenade along different ways to analyze them, we came across various representations. The large set of *snapshot* data from an intensive care unit allow to learn a set of decision functions at a central server, which could be applied in a streaming manner (Sect. 17.2.2). Learning classifiers from large sets of *value series* requires feature extraction. A unified framework for feature extraction from series allows to configure particular features from building blocks of temporal transformations and functions (Sect. 17.4). We stroll along other *feature extractions* as well. Counting how often an attribute in a temporal database changes its value can be a very effective feature (Sect. 17.3). *Events with intervals* are represented elegantly in predicate logic. A case study shows that robots can be controlled by learned logical rules (Sect. 17.6). Another logic-based approach investigates temporal cycles in economy (Sect. 17.5).

Learning algorithms along the way are Support Vector Machines (Sect. 17.2.1), (G)RDT from inductive logic programming (Sects. 17.5 and 17.6), and frequent set mining (Sect. 17.5), each requiring its appropriate representation. Being just a walk through diverse studies that dealt with temporal data, this chapter is far from being complete. The hope is—as is with a walk in the park—that some views might be inspiring for the own gardening and some just look nice.

Acknowledgements We gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG) during the SFB475, project A4 (1997–2009), the European Commission for the funding of the projects BLearn II (1992–1995), and MiningMart (2000–2003), and the DFG again for the support of the SFB876, project A1 (started 2011).

References

Ackermann, M. R., Lammersen, C., Märtens, M., Raupach, C., Sohler, C., & Swierkot, K. (2010). Streamkm++: a clustering algorithms for data streams. In G. Blelloch & D. Halperin (Eds.), *ALENEX* (pp. 173–187).

Agrawal, R., Imielinski, T., & Swami, A. (1993). Database mining: a performance perspective. *IEEE Transactions on Knowledge and Data Engineering*, *5*(6), 914–925.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large data bases. In *VLDB: Vol. 94. Proceedings of the 20th international conference on very large data bases* (pp. 478–499).

Allen, J. F. (1984). Towards a general theory of action and time. *Artificial Intelligence*, *23*, 123–154.

Borowski, M., Schettlinger, K., & Gather, U. (2009). Multivariate real-time signal extraction by a robust adaptive regression filter. *Communications in Statistics. Simulation and Computation*, *38*(2), 426–440.

Davy, M., Gretton, A., Doucet, A., & Rayner, P. (2002). Optimised support vector machines for nonstationary signal classification. *IEEE Transactions on Signal Processing*, *9*(12), 442–445.

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., & Keogh, E. (2008). Querying and mining of time series data: experimental comparison of representations and distance measures. In *Proceedings of the VLDB endowment* (Vol. 1, pp. 1542–1552).

Fried, R., & Gather, U. (2007). On rank tests for shift detection in time series. *Computational Statistics & Data Analysis*, *52*(1), 221–233.

Fried, R., Gather, U., Imhoff, M., & Bauer, M. (2000). Some statistical methods in intensive care online monitoring—a review. In R. W. Brause & E. Hanisch (Eds.), *ISMDA* (pp. 67–77). Berlin: Springer.

Geurts, P. (2001). Pattern extraction for time series classification. In *Proceedings of the 5th European conference on principles of data mining and knowledge discovery* (pp. 115–127). Berlin: Springer.

Heilemann, U., & Munch, H. (2001). *Classification of German business cycles using monthly data*. Technical report 08/2001, SFB 475, Universität Dortmund, Germany.

Höppner, F. (2002). Discovery of core episodes from sequences. In D. J. Hand, N. M. Adams, & R. J. Bolton (Eds.), *Pattern detection and discovery* (pp. 1–12). Berlin: Springer.

Joachims, T. (1999). Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods—support vector learning*, Cambridge: MIT Press.

Joachims, T. (2002). *Learning to classify text using support vector machines*. Norwell: Kluwer Academic.

Kaiser, M., & Dillmann, R. (1999). Introduction to skill learning. In K. Morik, V. Klingspor, & M. Kaiser (Eds.), *Making robots smarter—combining sensing and action through robot learning* (pp. 3–17). Norwell: Kluwer Academic.

Keogh, E., & Pazzani, M. (1998). An enhanced representation of time series which allows fast classification, clustering and relevance feedback. In *Proceedings of the 4th conference on knowledge discovery in databases* (pp. 239–241). Menlo Park: AAAI Press.

Keogh, E., & Pazzani, M. (2000). Scaling up dynamic time warping for data mining applications. In *Proceedings of the 6th conference on knowledge discovery and data mining* (pp. 285–289). Menlo Park: AAAI Press.

Kietz, J.-U., Vaduva, A., & Zucker, R. (2000). Mining mart: combining case-based-reasoning and multi-strategy learning into a framework to reuse KDD-application. In B. Michalski (Ed.), *Proceedings of the 5th international workshop on multistrategy learning*.

Kietz, J.-U., & Wrobel, S. (1992). Controlling the complexity of learning in logic through syntactic and task-oriented models. In S. Muggleton (Ed.), *Inductive logic programming* (pp. 335–360). San Diego: Academic Press.

Klingspor, V. (1994). GRDT: enhancing model-based learning for its application in robot navigation. In S. Wrobel (Ed.), *Proceedings of the 4th international workshop on inductive logic programming*.

Klingspor, V., & Morik, K. (1999). Learning understandable concepts for robot navigation. In K. Morik, V. Klingspor, & M. Kaiser (Eds.), *Making robots smarter—combining sensing and action through robot learning* (pp. 199–224). Norwell: Kluwer Academic.

Klinkenberg, R., & Joachims, T. (2000). Detecting concept drift with support vector machines. In P. Langley (Ed.), *Proceedings of the 7th international conference on machine learning* (pp. 487–494). San Mateo: Morgan Kaufmann.

Lanius, V., & Gather, U. (2010). Robust online signal extraction from multivariate time series. *Computational Statistics & Data Analysis*, *54*(4), 966–975.

Mannila, H., Toivonen, H., & Verkamo, A. (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, *1*(3), 259–290.

Mierswa, I., & Morik, K. (2005). Automatic feature extraction for classifying audio data. *Journal of Machine Learning Research*, *58*, 127–149.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). YALE: rapid prototyping for complex data mining tasks. In *Proceedings of the 12th conference on knowledge discovery and data mining* (pp. 935–940).

Morik, K., Wrobel, S., Kietz, J.-U., & Emde, W. (1993). *Knowledge acquisition and machine learning*. San Diego: Academic Press.

Morik, K. (2000). The representation race—preprocessing for handling time phenomena. In P. de Mántaras (Ed.), *Proceedings of the European conference on machine learning* (pp. 4–19). Berlin: Springer.

Morik, K., Brockhausen, P., & Joachims, T. (1999). Combining statistical learning with a knowledge-based approach—a case study in intensive care monitoring. In *Proceedings 16th* *international conference on machine learning*.

Morik, K., Joachims, T., Imhoff, M., Brockhausen, P., & Rüping, S. (2002). Integrating kernel methods into a knowledge-based approach to evidence-based medicine. In M. Schmitt et al. (Eds.), *Computational intelligence processing in medical diagnosis* (pp. 71–100). Heidelberg: Physica-Verlag.

Morik, K., & Rüping, S. (2002). A multistrategy approach to the classification of phases in business cycles. In *Proceedings of the European conference on machine learning* (pp. 307–318).

Morik, K., & Wessel, S. (1999). Incremental signal to symbol processing. In K. Morik, V. Klingspor, & M. Kaiser (Eds.), *Making robots smarter—combining sensing and action through robot learning* (pp. 185–198). Norwell: Kluwer Academic.

Morris, A. (1998). Algorithm-based decision making. In M. J. Tobin (Ed.), *Principles and practice of intensive care monitoring* (pp. 1355–1381). New York: McGraw-Hill.

Müller, K., Smola, A., Ratsch, G., Schölkopf, B., Kohlmorgen, J., & Vapnik, V. (1997). Predicting time series with support vector machines. In *Proceedings of the international conference on artificial neural networks*. Berlin: Springer.

Nunkesser, R., Fried, R., Schettlinger, K., & Gather, U. (2009). Online analysis of time series by the q_n estimator. *Computational Statistics & Data Analysis*, *53*(6), 2354–2362.

Oates, T., Firoiu, L., & Cohen, P. R. (2001). Using dynamic time warping to bootstrap HMM-based clustering of time series. In *Sequence learning—paradigms, algorithms, and applications* (pp. 35–52). Berlin: Springer.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. San Mateo: Morgan Kaufmann.

Rieger, A., & Klingspor, V. (1999). Program optimization for real-time performance. In K. Morik, V. Klingspor, & M. Kaiser (Eds.), *Making robots smarter—combining sensing and action through robot learning* (pp. 225–240). Norwell: Kluwer Academic.

Ritthoff, O., Klinkenberg, R., Fischer, S., & Mierswa, I. (2002). A hybrid approach to feature selection and generation using an evolutionary algorithm. In J. A. Bullinaria (Ed.), *Proceedings of the 2002 UK Workshop on Computational Intelligence (UKCI-02)* (pp. 147–154). Birmingham: University of Birmingham.

Rüping, S. (1999). *Zeitreihenprognose fur Warenwirtschaftssysteme unter Berücksichtigung asymmetrischer Kostenfunktionen*. Master thesis, Universität Dortmund.

Rüping, S. (2000) *mySVM-Manual*. University of Dortmund, Lehrstuhl Informatik 8. http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/.

Salton, G., & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.

Schlittgen, R., & Streitberg, B. H. J. (2001). *Zeitreihenanalyse*. Oldenburg

Sieben, W., & Gather, U. (2007). Classifying alarms in intensive care—analogy to hypothesis testing. In *Proceedings AIME* (pp. 130–138). Berlin: Springer.

Tzanetakis, G. (2002). *Manipulation, analysis and retrieval systems for audio signals*. Ph.D. Thesis, Computer Science Department, Princeton University.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Weihs, C., & Ligges, U. (2005). From local to global analysis of music time series. In S. Morik Boulicaut (Ed.), *Local pattern detection* (pp. 217–232). Berlin: Springer.

Wolpert, D., & Macready, W. (1997). No free lunch theorems for optimisation. *IEEE Transactions on Evolutionary Computation*, *1*, 67–82.

Wurst, M., Morik, K., & Mierswa, I. (2006). Localized alternative cluster ensembles for collaborative structuring. In S. Fürnkranz Scheffer (Ed.), *Proceedings of the European conference on machine learning* (pp. 485–496). Berlin: Springer.

Yi, B., Jagadish, H., & Faloutsos, C. (1998). Efficient retrieval of similar time series under time warping. In *Proceedings 14th conference on data engineering* (pp. 201–208).

Chapter 18
Correlation, Tail Dependence and Diversification

Dietmar Pfeifer

18.1 Introduction

What is frequently abbreviated as Solvency II is perhaps the most challenging legislative adventure in the European Union (besides Basel II/III for the banking sector) in the last decade. It is a fundamentally new, risk driven approach towards a harmonization of financial regulation for insurance and reinsurance companies writing business in the European Union. One of the major aims of the Solvency II framework is a customer protection limiting the yearly ruin probability of the company to at most 0.5 % by requiring sufficient economic capital. The calculation of this so called Solvency Capital Requirement (SCR) is based on a complicated mathematical and statistical framework derived from an economic balance sheet approach (for more details, see, e.g., Buckham et al. 2011; Cruz 2009; Doff 2007 or Sandström 2006). An essential aspect in the SCR calculation here is the notion of diversification, which aims at a reduction of the overall capital requirement by "distributing" risk in an appropriate way. There are several definitions and explanations of this term, some of which are presented in the sequel.

> "Although it is an old idea, the measurement and allocation of diversification in portfolios of asset and/or liability risks is a difficult problem, which has so far found many answers. The diversification effect of a portfolio of risks is the difference between the sum of the risk measures of stand-alone risks in the portfolio and the risk measure of all risks in the portfolio taken together, which is typically non-negative, at least for positive dependent risks."
>
> [Hürlimann (2009a, p. 325)]

> "Diversification arises when different activities complement each other, in the field of both return and risk. [...] The diversification effect is calculated by using correlation factors. Correlations are statistical measures assessing the extend to which events could occur simultaneously. [...] A correlation factor of 1 implies that certain events will always occur

D. Pfeifer (✉)
Institut für Mathematik, Carl von Ossietzky Universität Oldenburg, 26111 Oldenburg, Germany
e-mail: dietmar.pfeifer@uni-oldenburg.de

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_18, © Springer-Verlag Berlin Heidelberg 2013

> simultaneously. Hence, there is no diversification effect and two risks identically add up. Risk managers tend to say that such risks are perfectly correlated (i.e., they have a high correlation factor), meaning that these two risks do not actually diversify at all. A correlation factor of 0 implies that diversification effects are present and a certain diversification benefit holds."
>
> [Doff (2007, p. 167f.)]

> "By diversifiable we mean that if a risk category can be subdivided into risk classes and the risk charge of the total risk is not higher than the sum of the risk charges of each subrisk, then we have the effect of diversification. [...] This effect can be measured as the difference between the sum of several capital charges and the total capital charge when dependency between them is taken into account."
>
> [Sandström (2006, p. 188)]

> "In order to promote good risk management and align regulatory capital requirements with industry practices, the Solvency Capital Requirement should be determined as the economic capital to be held by insurance and reinsurance undertakings in order to ensure that ruin occurs no more often than once in every 200 cases. [...] That economic capital should be calculated on the basis of the true risk profile of those undertakings, taking account of the impact of possible risk-mitigation techniques, as well as diversification effects. [...] Diversification effects means the reduction in the risk exposure of insurance and reinsurance undertakings and groups related to the diversification of their business, resulting from the fact that the adverse outcome from one risk can be offset by a more favourable outcome from another risk, where those risks are not fully correlated. The Basic Solvency Capital Requirement shall comprise individual risk modules, which are aggregated [...] The correlation coefficients for the aggregation of the risk modules [...], shall result in an overall Solvency Capital Requirement [...] Where appropriate, diversification effects shall be taken into account in the design of each risk module."
>
> [Official Journal of the European Union (2009, (64) p. 7; (37) p. 24; Article 104, p. 52)]

One central idea that is common to all of these explanations is that a small, zero or even negative correlation between risks implies a diversification effect, while a large correlation or positive dependence implies the opposite. This is, however, largely based on a naïve understanding of the relationship between correlation and dependence which is not at all justified from a rigorous statistical point of view (see, e.g., Mari and Kotz 2001). This fact has also been emphasized by McNeil et al. (2005) in Chaps. 5 and 6 of their monograph, and in part also by Artzner et al. (1999).

A better way to tackle the understanding of a diversification effect is to replace the notion of correlation by the notion of copulas which describe the dependence structure between risks completely (see, e.g., Nelsen 2006, for a sophisticated survey). With respect to the "dangerousness" of joint risks, tail dependence is often used as a characteristic quantity (see, e.g., McNeil et al. 2005, Sect. 5.2.3). In case of a positive upper coefficient of tail dependence, it is likely that extreme events will occur more frequently simultaneously, just in the spirit of Doff's explanation of diversification above. This might suggest that risks with positive upper tail dependence are less exposed to diversification than those with zero upper tail dependence. However, a more sophisticated analysis shows that this is also not true in general.

The aim of this chapter is twofold:

Firstly, to show that the notion of correlation is completely disjoint from the notion of diversification under the risk measure VaR used in the Solvency II directive, i.e., we shall show that a state of no diversification between risks can be achieved with almost arbitrary positive and negative correlation coefficients, especially with the same marginal risk distributions.

And secondly, that a state of no diversification between risks can also be achieved with a zero tail dependence coefficient, or even worse, with a partial countermonotonic dependence structure, in particular for risks being lognormally distributed which is a basic assumption in the Pillar One standard model of Solvency II.

18.2 A Short Review of Risk Measures

In this section, we shall only focus on risk measures for non-negative risks since these are the essential quantities in insurance, and are also the fundamentals of the SCR calculation under Solvency II. We follow a simplified setup as in Sandström (2006), Sect. 7.4 which is formally slightly different from the approach in Artzner et al. (1999) or McNeil et al. (2005, Chap. 6).

Definition 18.1 Let $\mathcal{X}$ Let be a suitable set of non-negative random variables X on a probability space $(\Omega, \mathcal{A}, P)$. A risk measure R on $\mathcal{X}$ is a mapping $\mathcal{X} \to \mathbb{R}^+$ with the following properties:

$$P^X = P^Y \Rightarrow R(X) = R(Y) \quad \forall X, Y \in \mathcal{X}, \tag{18.1}$$

i.e., the risk measure depends only on the distribution of the risk X;

$$R(cX) = cR(X) \quad \forall X \in \mathcal{X} \quad \forall c \geq 0,$$

i.e., the risk measure is *scale-invariant*;

$$R(X + c) = R(X) + c \quad \forall X \in \mathcal{X} \quad \forall c \geq 0, \tag{18.2}$$

i.e., the risk measure is *translation-invariant*;

$$R(X) \leq R(Y) \quad \forall X, Y \in \mathcal{X}, \; X \leq Y, \tag{18.3}$$

i.e., the risk measure is *monotone*.

The risk measure is called *coherent*, if it additionally has the subadditivity property:

$$R(X + Y) \leq R(X) + R(Y) \quad \forall X, Y \in \mathcal{X}. \tag{18.4}$$

This last property is the crucial point: it induces a diversification effect for *arbitrary* non-negative risks $X_1, \ldots, X_n$ (dependent or not) since it follows by induction that coherent risk measures have the property

$$R\left(\sum_{k=1}^{n} X_k\right) \leq \sum_{k=1}^{n} R(X_k) \quad \forall n \in \mathbb{N}.$$

In what follows we shall use the term "(risk) *concentration* effect" as opposite to "*diversification* effect", characterized by the converse inequality

$$\exists (X,Y) \in \mathcal{X} \times \mathcal{X} : R(X+Y) > R(X) + R(Y).$$

Example 18.1 The popular standard deviation principle *SDP* which is sometimes used for tariffing in insurance is defined as

$$\mathrm{SDP}(X) = E(X) + \gamma\sqrt{\mathrm{Var}(X)} \quad \text{for a fixed } \gamma > 0 \text{ and } X \in \mathcal{X} = \mathcal{L}^2_+(\Omega, \mathcal{A}, P),$$

the set of non-negative square-integrable random variables on $(\Omega, \mathcal{A}, P)$. Obviously, SDP fulfils the properties (18.1) to (18.2) and (18.4); the latter because of

$$\begin{aligned} \mathrm{SDP}(X+Y) &= E(X) + E(Y) \\ &\qquad + \gamma\sqrt{\mathrm{Var}(X) + \mathrm{Var}(Y) + 2\rho(X,Y)\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}} \\ &\le E(X) + E(Y) + \gamma\sqrt{\mathrm{Var}(X) + \mathrm{Var}(Y) + 2\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}} \\ &= E(X) + E(Y) + \gamma\sqrt{\left(\sqrt{\mathrm{Var}(X)} + \sqrt{\mathrm{Var}(Y)}\right)^2} \\ &= \mathrm{SDP}(X) + \mathrm{SDP}(Y) \end{aligned} \tag{18.5}$$

for all $X, Y \in \mathcal{X}$. Here $\rho(X,Y) = \mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$ denotes the correlation between X and Y. However, *SDP* does in general not fulfil property (18.3) and is hence not a proper risk measure, as can be seen as follows: Let Z be a random variable binomially distributed over $\{0,1\}$ with $P(Z=1) = 1 - P(Z=0) = p$, and $1/(1+\gamma^2) < p < 1$. Consider $X := 2Z$ and $Y := 1 + Z$. Then $X \le Y$, but $R(X) = 2p + 2\gamma\sqrt{p(1-p)} > 1 + p + \gamma\sqrt{p(1-p)} = R(Y)$.

Example 18.2 The risk measure used in Basel II/III and Solvency II is the Value-at-Risk VaR, being defined as a (typically high) quantile of the risk distribution:

$$\mathrm{VaR}_\alpha(X) := Q_X(1-\alpha) \quad \text{for } X \in \mathcal{X} \text{ and } 0 < \alpha < 1,$$

where Q_X denotes the quantile function

$$Q_X(u) := \inf\left\{x \in \mathbb{R} \mid P(X \le x) \ge u\right\} \quad \text{for } 0 < u < 1.$$

Value-at-Risk is a proper risk measure, but not coherent in general. This topic will be discussed in more detail in the next section (for a more general discussion, see e.g., McNeil et al. 2005, Sect. 6.1.2).

The "smallest" coherent risk measure above VaR is the expected shortfall (ES), which is in general defined as

$$\mathrm{ES}_\alpha(X) := \frac{1}{\alpha}\left\{E(X\cdot \mathbb{I}_{\{X \ge \mathrm{VaR}_\alpha(X)\}}) + \mathrm{VaR}_\alpha(X)\left[\alpha - P\left(X \ge \mathrm{VaR}_\alpha(X)\right)\right]\right\}$$

for $X \in \mathcal{X}$ and $0 < \alpha < 1$, where $\mathbb{I}_A$ denotes the indicator random variable of some event (measurable set) A. In case that $P(X \geq \mathrm{VaR}_\alpha(X)) = \alpha$, this formula simplifies to

$$\mathrm{ES}_\alpha(X) = E\big(X \mid X \geq \mathrm{VaR}_\alpha(X)\big) = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_u(X)\,du$$

(see McNeil et al. 2005, Definition 2.15 and Remark 2.17); note that the role of α and $1-\alpha$ are interchanged there). A more thorough discussion on the relationship between VaR and ES (and other coherent risk measures) in connexion with Wang's distortion measures can be found in Hürlimann (2004). Expected shortfall is the risk measure which is used in the Swiss Solvency Test (SST), see, e.g., Sandström (2006, Sect. 6.8) or Cruz (2009, Chap. 17).

18.3 A Short Review of Copulas

The copula approach allows for a separate treatment of the margins of joint risks and the dependence structure between them. The name "copula" goes back to Abe Sklar in 1959 who used it as a function which couples a joint distribution function with its univariate margins. For an extensive survey, see, e.g., Nelsen (2006).

Definition 18.2 A copula (in n dimensions) is a function C defined on the unit cube $[0,1]^n$ with the following properties:

1. the range of C is the unit interval $[0,1]$;
2. $C(\mathbf{u})$ is zero for all $\mathbf{u} = (u_1,\dots,u_n)$ in $[0,1]^n$ for which at least one coordinate is zero;
3. $C(\mathbf{u}) = u_k$ if all coordinates of $\mathbf{u}$ are 1 except the k-th one;
4. C is n-increasing in the sense that for every $\mathbf{a} \leq \mathbf{b}$ in $[0,1]^n$ the volume assigned by C to the subinterval $[\mathbf{a},\mathbf{b}] = [a_1,b_1] \times \cdots \times [a_n,b_n]$ is nonnegative.

A copula can alternatively be characterized as a multivariate distribution function with univariate marginal distribution functions that belong to a continuous uniform distribution over the unit interval $[0,1]$.

It can be shown that every copula is bounded by the so-called Fréchet–Hoeffding bounds, i.e.,

$$\begin{aligned} C_*(\mathbf{u}) &:= \max(u_1 + \cdots + u_n - n + 1, 0) \leq C(u_1,\dots,u_n) \\ &\leq C^*(\mathbf{u}) := \min(u_1,\dots,u_n). \end{aligned}$$

The upper Fréchet–Hoeffding bound C^* is a copula itself for any dimension; however, the lower Fréchet–Hoeffding bound C_* is a copula in two dimensions only. If X is any real random variable, then the random vector $\mathbf{X} = (X, X, \dots, X)$ with n components possesses the upper Fréchet–Hoeffding bound C^* as copula, while the random vector $\mathbf{X} = (X, -X)$ with two components possesses the lower Fréchet–Hoeffding bound C_* as copula. Random variables who have C^* or C_*, respectively, as copula are also called *comonotone* or *countermonotone*, respectively. An important and well-studied copula is the independence copula, given by $C(\mathbf{u}) = \prod_{i=1}^{n} u_i$.

The following theorem due to Sklar justifies the role of copulas as dependence functions:

Proposition 18.1 *Let H denote some n-dimensional distribution function with marginal distribution functions $F_1, \dots, F_n$. Then there exists a copula C such that for all real $(x_1, \dots, x_n)$,*

$$H(x_1, \dots, x_n) = C\bigl(F_1(x_1), \dots, F_n(x_n)\bigr).$$

If all the marginal distribution functions are continuous, then the copula is unique. Moreover, the converse of the above statement is also true. If we denote by $F_1^{-1}, \dots, F_n^{-1}$ the generalized inverses of the marginal distribution functions (or quantile functions), then for every $(u_1, \dots, u_n)$ in the unit cube,

$$C(u_1, \dots, u_n) = H\bigl(F_1^{-1}(u_1), \dots, F_n^{-1}(u_n)\bigr).$$

For a proof, see (Nelsen 2006, Theorem 2.10.9) and the references given therein. The above theorem shows that copulas remain invariant under strictly monotone transformations of the same kind of the underlying random variables (either increasing or decreasing).

The following result shows the relationship between correlation and copulas.

Proposition 18.2 *Let (X, Y) be a bivariate random vector with a copula C and marginal distribution functions F and G such that $E(|X|) < \infty$, $E(|Y|) < \infty$ and $E(|XY|) < \infty$. Then the covariance between X and Y can be expressed in the following way:*

$$\operatorname{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \bigl[C\bigl(F(x), G(y)\bigr) - F(x)G(y)\bigr]\, dx\, dy.$$

For a proof see, e.g., McNeil et al. (2005, Lemma 5.24).

18.4 Correlation and Diversification

Before going into deeper details, we start with an illustrative example showing essentially that risk concentration under VaR can occur with almost all positive and negative correlation coefficients, even if the marginal distributions remain fixed. According to the Solvency II standard, we choose $\alpha = 0.005$ for simplicity here, but the example can be reformulated accordingly with any other value of $0 < \alpha < 1$.

Table 18.1 Joint distribution of risks

$P(X=x,Y=y)$		x			$P(Y=y)$	$P(Y\leq y)$
		0	50	100		
y	0	β	$0.440-\beta$	0.000	0.440	0.440
	40	$0.554-\beta$	β	0.001	0.555	**0.995**
	50	0.000	0.001	0.004	0.005	1.000
$P(X=x)$		0.554	0.441	0.005		
$P(X\leq x)$		0.554	**0.995**	1.000		

Table 18.2 Moments and correlations

$E(X)$	$E(Y)$	$\sigma(X)$	$\sigma(Y)$	$\rho(\beta)=\rho(X,Y)$
22.550	22.450	25.377	19.912	$-0.9494+3.9579\beta$

Table 18.3 Distribution of aggregate risk

s	0	40	50	90	100	140	150
$P(S=s)$	β	$0.554-\beta$	$0.440-\beta$	β	0.001	0.001	0.004
$P(S\leq s)$	β	0.554	$0.994-\beta$	0.994	**0.995**	0.996	1.000

Example 18.3 Let the joint distribution of the non-negative risks X and Y be given by Table 18.1, with $0\leq\beta\leq 0.440$, giving $\mathrm{VaR}_\alpha(X)=50$, $\mathrm{VaR}_\alpha(Y)=40$.

For the moments of X and Y, we obtain the values in Table 18.2 (with σ denoting the standard deviation). This shows that the range of possible risk correlations is the interval $[-0.9494; 0.7921]$, with a zero correlation being attained for $\beta=0.2399$.

Table 18.3 shows the distribution of the aggregated risk $S=X+Y$.

We thus obtain a risk concentration due to $\mathrm{VaR}_\alpha(S)=100>90=\mathrm{VaR}_\alpha(X)+\mathrm{VaR}_\alpha(Y)$, independent of the parameter β and hence also independent of the possible correlations between X and Y.

A closer look to the joint distribution of X and Y shows that the reason for this perhaps unexpected result is the fact that although one can have a "diversification effect" in the central body of the distribution, where a fraction of little less than $1-\alpha$ of the risk pairs are located, the essential "concentration effect", however, is caused by a joint occurrence of very high losses, with a fraction of α of all risk pairs.

The following result is related to the consideration of "worst VaR scenarios" as in McNeil et al. (2005, Sect. 6.2).

Proposition 18.3 *Let X and Y be non-negative risks with cumulative distribution functions F_X and F_Y, respectively, which are continuous and strictly increasing on* *their support. Denote, for a fixed* $\alpha \in (0,1)$,

$$Q^*(\alpha,\delta) := \min\big\{Q_X(u) + Q_Y(2-\alpha-\delta-u) \mid 1-\alpha-\delta \le u \le 1\big\}$$
$$\textit{for } 0 \le \delta < 1-\alpha.$$

Then there exists a sufficiently small $\epsilon \in (0, 1-\alpha)$ *with the property*

$$Q^*(\alpha,\epsilon) > Q_X(1-\alpha) + Q_Y(1-\alpha) = \mathrm{VaR}_\alpha(X) + \mathrm{VaR}_\alpha(Y). \tag{18.6}$$

Assume further that the random vector (U, V) *has a copula C as joint distribution function with the properties*

$$\begin{aligned} V < 1-\alpha-\epsilon \;&\Longleftrightarrow\; U < 1-\alpha-\epsilon \quad \textit{and}\\ V = 2-\alpha-\epsilon-U \;&\Longleftrightarrow\; U \ge 1-\alpha-\epsilon. \end{aligned} \tag{18.7}$$

If we define

$$X^* := Q_X(U), \quad Y^* := Q_Y(V), \quad S^* := X^* + Y^*,$$

then the random vector (X^*, Y^*) *has the same marginal distributions as* (X, Y), *and it holds*

$$\mathrm{VaR}_\alpha\big(X^*+Y^*\big) \ge Q^*(\alpha,\epsilon) > \mathrm{VaR}_\alpha\big(X^*\big) + \mathrm{VaR}_\alpha\big(Y^*\big) = \mathrm{VaR}_\alpha(X) + \mathrm{VaR}_\alpha(Y),$$

i.e., there is a risk concentration effect. Moreover, under the assumption (18.7), *the correlation* $\rho(X^*, Y^*)$ *is minimal if* $V = 1-\alpha-\epsilon-U$ *for* $U < 1-\alpha-\epsilon$ *(lower extremal copula* $\underline{C}$*) and maximal if* $V = U$ *for* $U < 1-\alpha-\epsilon$ *(upper extremal copula* $\overline{C}$*).*

Proof By assumption, the (non-negative) quantile functions Q_X and Q_Y are continuous and strictly increasing over the interval [0, 1] (with a possibly infinite value at the point 1), so that

$$\begin{aligned} &Q^*(\alpha,0)\\ &\quad = \min\big\{Q_X(u) + Q_Y(2-\alpha-u) \mid 1-\alpha \le u \le 1\big\} > Q_X(1-\alpha) + Q_Y(1-\alpha), \end{aligned}$$

the minimum being actually attained. Since by the continuity assumptions above, $Q^*(\alpha,\epsilon)$ is continuous in ϵ and decreasing when ϵ is increasing, relation (18.6) follows.

The copula construction above now implies that

$$\begin{aligned} &P\big(S^* \le s\big) \le 1-\alpha-\epsilon\\ &\quad \text{for } s \le Q_X(1-\alpha-\epsilon) + Q_Y(1-\alpha-\epsilon) = \mathrm{VaR}_{\alpha+\epsilon}(X) + \mathrm{VaR}_{\alpha+\epsilon}(Y),\\ &P\big(S^* \le s\big) = 1-\alpha-\epsilon \quad \text{for } \mathrm{VaR}_{\alpha+\epsilon}(X) + \mathrm{VaR}_{\alpha+\epsilon}(Y) < s < Q^*(\alpha,\epsilon),\\ &P\big(S^* \le s\big) \ge 1-\alpha-\epsilon \quad \text{for } s \ge Q^*(\alpha,\epsilon). \end{aligned} \tag{18.8}$$

Table 18.4 Examples of risk measures and correlations for various values of σ

σ	$\mathrm{VaR}_\alpha(X)$ $= \mathrm{VaR}_\alpha(Y)$	$\mathrm{VaR}_\alpha(X)$ $+ \mathrm{VaR}_\alpha(Y)$	$\mathrm{VaR}_\alpha(X^* + Y^*)$	$\rho_{\min}(X^*, Y^*)$	$\rho_{\max}(X^*, Y^*)$
0.1	1.2873	2.5746	2.6205	−0.8719	0.9976
0.2	1.6408	3.2816	3.3994	−0.8212	0.9969
0.3	2.0704	4.1408	4.3661	−0.7503	0.9951
0.4	2.5866	5.1732	5.5520	−0.6620	0.9920
0.5	3.1992	6.3984	6.9901	−0.5598	0.9873
0.6	3.9177	7.8354	8.7134	−0.4480	0.9802
0.7	4.7497	9.4994	10.7537	−0.3310	0.9700
0.8	5.7011	11.4022	13.1401	−0.2136	0.9556
0.9	6.7750	13.5500	15.8969	−0.1002	0.9362
1.0	7.9712	15.9424	19.0412	0.0050	0.9108
1.5	15.4675	30.9350	40.4257	0.3127	0.6839
2.0	23.3748	46.7496	66.8923	0.2723	0.3794
2.5	27.5107	55.0214	86.2673	0.1399	0.1637
3.0	25.2162	50.4324	86.7034	0.0565	0.0611

Relation (18.8) in turn implies that $\mathrm{VaR}_\alpha(S^*) = \mathrm{VaR}_\alpha(X^* + Y^*) \geq Q^*(\alpha, \epsilon)$ which proves the first part of Proposition 18.3, due to relation (18.6).

The remainder part follows from Theorem 5.25 in McNeil et al. (2005) when looking at the conditional distribution of (X^*, Y^*) given the event $\{U < 1 - \alpha - \epsilon\}$. □

Note that both types of copulas that provide the extreme values for the correlations, $\underline{C}$ and $\overline{C}$, are of the type "shuffles of M", see Nelsen (2006, Sect. 3.2.3).

In the following example, we shall show some consequences of Proposition 18.3 in the case of lognormally distributed risks, which are of special importance for Pillar One under Solvency II, see, e.g., Hürlimann (2009a, 2009b).

Example 18.4 To keep things simple and comparable with Solvency II specifications, we shall assume that X and Y follow the same lognormal distribution $\mathcal{LN}(\mu, \sigma)$ with $\mu \in \mathbb{R}$, $\sigma > 0$ and $E(X) = E(Y) = 1$ which corresponds to the case $\mu = -\sigma^2/2$. Table 18.4 shows all relevant numerical results for the extreme copulas $\underline{C}$ and $\overline{C}$ in Proposition 18.3, especially the maximal range of correlations induced by them. According to the Solvency II standard, we choose $\alpha = 0.005$ (and $\epsilon = 0.001$, which will be sufficient here).

Note that the bottom graph in Fig. 18.1 resembles the graph in Fig. 5.8 in McNeil et al. (2005).

The graph in Fig. 18.2 shows parts of the two cumulative distribution functions under the extreme copulas $\underline{C}$ and $\overline{C}$ for $S^* := X^* + Y^*$ in the case $\sigma = 1$. Note that especially for smaller values of σ (which is typical for the calculation of the SCR in

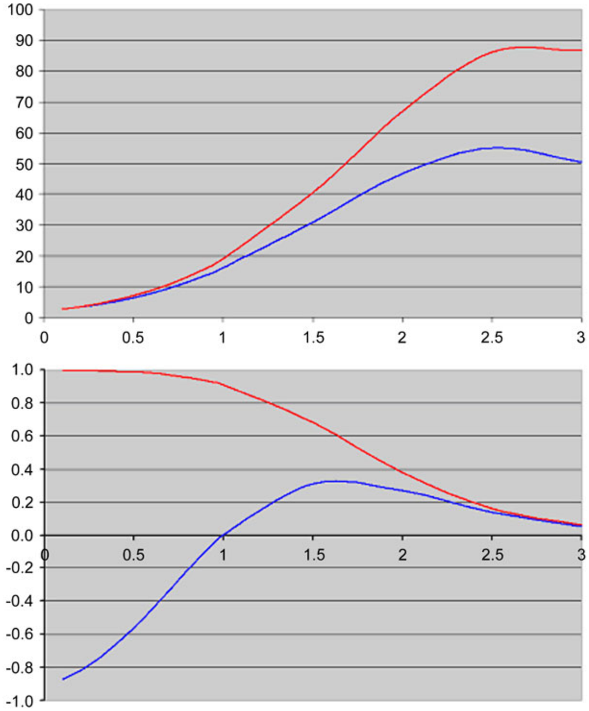


Fig. 18.1 *Top*: Graph of $\mathrm{VaR}_\alpha(X^* + Y^*)$ (*red*) and $\mathrm{VaR}_\alpha(X) + \mathrm{VaR}_\alpha(Y)$ (*blue*) as functions of σ; *bottom*: graph of $\rho_{\max}(X^*, Y^*)$ (*red*) and $\rho_{\min}(X^*, Y^*)$ (*blue*) as functions of σ

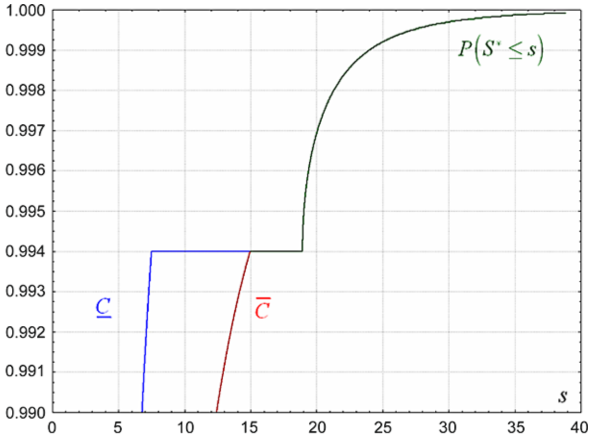


Fig. 18.2 Graph of cumulative distribution functions for extreme copulas for $S^* = X^* + Y^*$ with $\sigma = 1$

the non-life risk module of Solvency II) the range of possible negative and positive correlations between the risks is quite large, with the same significant discrepancy between the Value at Risk of the aggregated risks and the sum of individual Values at Risk.

Note also that any correlation value ρ of $\rho(X^*, Y^*)$ between $\rho_{\min}(X^*, Y^*)$ and $\rho_{\max}(X^*, Y^*)$ can be achieved by a proper mixture of the extreme copulas $\underline{C}$ and $\overline{C}$

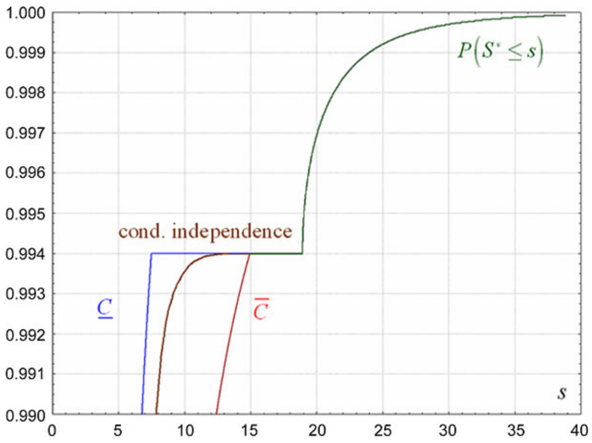


Fig. 18.3 Graph of cumulative distribution functions for extreme copulas and conditional independence

namely for the copula

$$C(p) = \lambda\underline{C} + (1-\lambda)\overline{C} \quad \text{with } \lambda = \frac{\rho_{\max}(X^*,Y^*) - \rho}{\rho_{\max}(X^*,Y^*) - \rho_{\min}(X^*,Y^*)}.$$

This is a direct consequence from Proposition 18.2, for example.

There are, of course, also other possibilities to achieve appropriate intermediary values for the correlation, for instance if U and V are conditionally independent given the event $\{U < 1-\alpha-\epsilon\}$. The graph in Fig. 18.3 adds a part of the cumulative distribution function of $S^* := X^* + Y^*$ for this case to the graph in Fig. 18.2. The correlation between X^* and Y^* is here given by $\rho(X^*,Y^*) = 0.3132$.

18.5 Tail Dependence and Diversification

As in the case of a large positive correlation between risks, it might be intuitively tempting to assume that a positive upper tail dependence would have a positive impact on risk concentration, too. But this is not true here either. In this section, we shall show that a risk concentration effect can occur with and without tail dependence, while the marginal distributions remain unchanged.

Note first that the copula construction of Proposition 18.3 implies no upper tail dependence since, by the continuity assumption for the marginal distributions made there (see, e.g., McNeil et al. 2005, Sect. 5.2.3),

$$\lambda_u = \lim_{u \uparrow 1} \frac{P(U > u, V > u)}{1-u} = 0$$

because for $1-(\alpha+\epsilon)/2 < u \le 1$, we have $2-\alpha-\epsilon-u < u$ and hence, for these u,

$$\begin{aligned} P(U > u, V > u) &= P(U > u, 2-\alpha-\epsilon-U > u) \\ &= P(u < U < 2-\alpha-\epsilon-u) = P(\emptyset) = 0. \end{aligned}$$

The following proposition shows that we can incorporate an upper tail dependence in the construction of Proposition 18.3 without essentially loosing the central result.

Proposition 18.4 *Assume that the conditions of Proposition* 18.3 *hold, with the following modification of the copula construction in* (18.7):

$$V < 1-\alpha-\epsilon \iff U < 1-\alpha-\epsilon \quad \textit{and}$$
$$V = \begin{cases} 2-\alpha-\epsilon-\gamma-U : 1-\alpha-\epsilon \le U < 1-\gamma \\ U : 1-\gamma \le U \le 1 \end{cases}$$

with some non-negative $\gamma < \alpha$. Then, for sufficiently small ϵ and γ, we still have

$$\begin{aligned} &\min\bigl\{Q_X(u) + Q_Y(2-\alpha-\epsilon-\gamma-u) \mid 1-\alpha-\epsilon \le u \le 1-\gamma\bigr\} \\ &\quad > Q_X(1-\alpha) + Q_Y(1-\alpha) \end{aligned}$$

and hence again a risk concentration effect, i.e., $\mathrm{VaR}_\alpha(X^* + Y^*) > \mathrm{VaR}_\alpha(X) + \mathrm{VaR}_\alpha(Y)$. *Moreover, under this copula construction, the correlation $\rho(X^*, Y^*)$ is again minimal if $V = 1-\alpha-\epsilon-U$ for $U < 1-\alpha-\epsilon$ (lower extremal copula $\underline{C}_\gamma$) and maximal if $V = U$ for $U < 1-\alpha-\epsilon$ (upper extremal copula $\overline{C}_\gamma$). Further, the risks are in all cases upper tail dependent with*

$$\lambda_u = \lim_{u \uparrow 1} \frac{P(U > u, V > u)}{1-u} = 1. \tag{18.9}$$

Proof The first two parts follow along the lines of the proof of Proposition 18.3. For the last part, observe that we have $U = V \iff 1-\gamma \le U \le 1$ which implies $P(U > u, V > u) = 1-u$ for $1-\gamma \le u \le 1$ and hence (18.9). □

Example 18.5 If we choose $\gamma = 0.0005$, we get the extension of the results in Example 18.4, under the same initial conditions (see Table 18.5). The graph in Fig. 18.4 shows an extension of the graph in Fig. 18.3 for the extreme copulas $\underline{C}_\gamma$ and $\overline{C}_\gamma$ for $S^* := X^* + Y^*$ in the case $\sigma = 1$. The case $\gamma > 0$ corresponds to upper tail dependence, the case $\gamma = 0$ correspond to the former situation with no upper tail dependence.

18.6 Conclusions

As the preceding analysis has shown, neither the notion of correlation nor the notion of tail dependence as such has in general a direct impact on diversification under the

Table 18.5 Examples of risk measures and correlations for various values of σ

σ	$\mathrm{VaR}_\alpha(X)$ $= \mathrm{VaR}_\alpha(Y)$	$\mathrm{VaR}_\alpha(X)$ $+ \mathrm{VaR}_\alpha(Y)$	$\mathrm{VaR}_\alpha(X^* + Y^*)$	$\rho_{\min}(X^*, Y^*)$	$\rho_{\max}(X^*, Y^*)$
0.1	1.2873	2.5746	2.6134	−0.8710	0.9993
0.2	1.6408	3.2816	3.3811	−0.8193	0.9988
0.3	2.0704	4.1408	4.3308	−0.7471	0.9981
0.4	2.5866	5.1732	5.4923	−0.6568	0.9969
0.5	3.1992	6.3984	6.8962	−0.5515	0.9953
0.6	3.9177	7.8354	8.5730	−0.4349	0.9929
0.7	4.7497	9.4994	10.5516	−0.3107	0.9974
0.8	5.7011	11.4022	12.8581	−0.1830	0.9964
0.9	6.7750	13.5500	15.5133	−0.0553	0.9951
1.0	7.9712	15.9424	18.5310	0.0691	0.9744
1.5	15.4675	30.9350	38.8061	0.5658	0.9366
2.0	23.3748	46.7496	63.3300	0.8154	0.9224
2.5	27.5107	55.0214	80.5429	0.9185	0.9423
3.0	25.2162	50.4324	79.8272	0.9636	0.9909

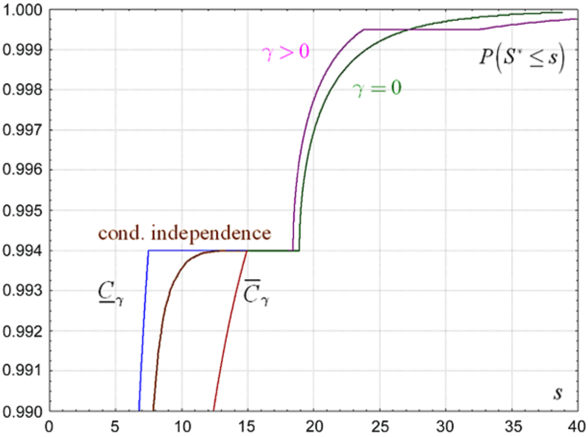


Fig. 18.4 Graph of cumulative distribution functions for extreme copulas and conditional independence, with and without upper tail dependence

risk measure Value at Risk. This means that any attempt to implement such concepts into a simple Pillar One standard model under Solvency II for the purpose of a reduction of the Solvency Capital Requirement in case of a diversification effect cannot be justified by mathematical reasoning. We can perhaps summarize the consequences of this insight in a slight modification of a statement in McNeil et al. (2005, p. 205):

> The concept of diversification is meaningless unless applied in the context of a well-defined joint model. Any interpretation of diversification in the absence of such a model should be avoided.

References

Artzner, P., Delbaen, F., Eber, J. M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*, 203–228.

Buckham, D., Wahl, J., & Rose, S. (2011). *Executives guide to solvency II*. New York: Wiley.

Cruz, M. (2009). *The solvency II handbook. Developing ERM frameworks in insurance and reinsurance companies*. London: Risk Books.

Doff, R. (2007). *Risk management for insurers. Risk control, economic capital and solvency II*. London: Risk Books.

Hürlimann, W. (2004). Distortion risk measures and economic capital. *North American Actuarial Journal*, *8*, 86–95.

Hürlimann, W. (2009a). Optimisation of the non-life insurance risk diversification in solvency II. In M. Cruz (Ed.), *The solvency II handbook. Developing ERM frameworks in insurance and reinsurance companies* (pp. 325–347). London: Risk Books. Chap. 12.

Hürlimann, W. (2009b). On the non-life solvency II model. In M. Cruz (Ed.), *The solvency II handbook. Developing ERM frameworks in insurance and reinsurance companies* (pp. 349–370). London: Risk Books. Chap. 13.

Mari, D. D., & Kotz, S. (2001). *Correlation and dependence*. London: Imperial College Press.

McNeil, A. J., Frey, R., & Embrechts, P. (2005). *Quantitative risk management—concepts, techniques, tools*. Princeton: Princeton University Press.

Nelsen, R. B. (2006). *An introduction to copulas*. New York: Springer.

Official Journal of the European Union (2009). Directive 2009/138/EC of the European parliament and of the council of 25 November 2009 on the taking-up and pursuit of the business of insurance and reinsurance (Solvency II). http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:335:FULL:EN:PDF

Sandström, A. (2006). *Solvency. Models, assessment and regulation*. London/Boca Raton: Chapman & Hall/CRC Press.

Chapter 19
Evidence for Alternative Hypotheses

Stephan Morgenthaler and Robert G. Staudte

19.1 Introduction

Statisticians are trained to avoid “lying with statistics,” that is, to avoid deceiving others and themselves about what the data say about questions or hypotheses. At the most fundamental level, they are battling against the power of *one* number to influence thinking, rather than two numbers. Telling someone that “smoking doubles the risk of lung cancer” is a powerful message, likely to be accepted as a fact. But reporting the “two”, with a standard error, or reporting a confidence interval for the relative risk is likely to have far less impact. It is as if there were less reliability in the message with the greater information, no doubt because the second number reminds us of the imprecision in the first. Statisticians are not immune from this human fallibility. We often quote a p-value against a null hypothesis, or a posterior probability for a hypothesis or a likelihood ratio for comparing two hypotheses, as if they were important numerical facts, to be taken at face-value, without further question. Evans (2000) in his comments on Royall (2000), makes the same point: “Some quantification concerning the uncertainty inherent in what the likelihood ratio is saying seems to be a part of any acceptable theory of statistical inference. In other words, such a quantification is part of the summary of statistical evidence.” We agree with Evans and thus require that any measure of statistical evidence be a statistic reported with a standard deviation or other measure of uncertainty.

Another example of an incomplete message occurs frequently in the meta-analytic literature. Results are derived for the case of known weights, and then es-

S. Morgenthaler (✉)
École Polytechnique Fédérale de Lausanne, EPFL–FSB–STAP, Station 8, 1015 Lausanne, Switzerland
e-mail: stephan.morgenthaler@epfl.ch

R.G. Staudte
Department of Mathematics and Statistics, La Trobe University, Melbourne, Victoria 3086, Australia
e-mail: r.staudte@latrobe.edu.au

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_19, © Springer-Verlag Berlin Heidelberg 2013

timates of the weights are substituted in the ensuing formulae, as if no theory were needed to account for the second estimation. This works for very large sample sizes, but not for those usually encountered in practice and results in optimistically small confidence intervals, inflated coverages and many published false claims, Malzahn et al. (2000), e.g., thus, we require that any measure of evidence found for individual studies of the same effect should be easy to combine to obtain an overall evidence for this effect, and the combination of evidence must be based on a sound theory (see the contribution by Wellmann, Chap. 22, for a discussion of meta analysis). In the remainder of this section, we motivate and define statistical evidence on our preferred calibration scale in which one function, called the Key Inferential Function, contains all the information required for inference.

For the sake of simplicity of presentation, we restrict attention to one-sided alternatives $\theta > \theta_0$ to the null hypothesis $\theta = \theta_0$ (or $\theta \leq \theta_0$); evidence for two-sided alternatives is presented in detail in Sect. 17.4, p. 134 of Kulinskaya et al. (2008). Our third requirement is that the expected evidence in favor of $\theta > \theta_0$ should be increasing with θ and have value 0 at $\theta = \theta_0$.

Fourth, if the parameter of interest θ is estimable by a $\hat{\theta}_n$ based on n observations, with standard error SE$[\hat{\theta}_n]$ of order $1/\sqrt{n}$, then the evidence for an alternative hypothesis $\theta > \theta_0$ should grow at the rate $\sqrt{n}$. This means it will require 9 times as much work to obtain 3 times as much evidence for an alternative hypothesis.

Fifth, evidence should be replicable in the sense that if an experimenter obtains a certain amount of evidence for a hypothesis, then an independent repetition of the experiment should lead to a similar result, up to sampling error. While this is true for the p-value under the null hypothesis, when the null hypothesis does not hold the variation under repetition may come as a surprise due to the highly skewed distribution of the p-value. These five motivating factors lead us to illustrate what is achievable for the simplest possible model in the next section, a model that forms the basis for all that follows.

Prototypical Example We will now consider an example that is often discussed in elementary statistics courses. In this example, each experiment produces an independent realization of a random variable $X \sim \mathbf{N}(\mu, \sigma_0^2)$, where σ_0^2 is known. This is the prototypical normal translation model, which we would like to use as a "universal model" for other testing problems. We want to quantify the evidence based on n experiments against the null hypothesis $\mu = \mu_0$ and in favor of the alternative $\mu > \mu_0$. Letting $\bar{X}_n$ denote the sample mean, the usual test statistic is $S_n = \bar{X}_n - \mu_0$. The corresponding evidence is the standardized version of S_n, that is, $T_n = \sqrt{n}(\bar{X}_n - \mu_0)/\sigma_0$, which is distributed as $\mathbf{N}(\sqrt{n}(\mu - \mu_0)/\sigma_0, 1)$. The way an evidence is constructed means that the expected evidence is the function of the parameters that carries all the information. In the normal shift model, this is $\sqrt{n}(\mu - \mu_0)/\sigma_0$. Because T_n has variance one, evidence can be reported as $T_n \pm 1$, indicating that it has error, with the subtext that this standard error means the same thing to everyone, because all students of statistics recognize a standard normal distribution. This standard error of 1 also becomes the unit for a calibration scale for evidence: if one observes $T_n = 3$, one knows that one has observed a result 3 times its own standard error. If one obtains $T_n = -2$, one has evidence $+2$ for the opposite alternative hypothesis $\mu < \mu_0$, again with standard normal error.

The statistical evidence $T_n = \sqrt{n}(\bar{X}_n - \mu_0)/\sigma_0$ is monotone increasing in $\sqrt{n}$ for each fixed μ; and, for each fixed sample size n, the expected evidence grows from 0 as $\mu > \mu_0$ increases. This evidence is also replicable in the sense that given $T_n = t$ and an independent $T_n^* \sim \mathbf{N}(\sqrt{n}(\mu - \mu_0)/\sigma_0, 1)$ the optimal predictor $\mathrm{E}[T_n^* \mid T_n = t]$ is simply t, and this estimator of the expected evidence has standard error 1.

When combining evidence from independent studies, given $T_{n_1} \sim N(\tau_1, 1)$ and $T_{n_2} \sim N(\tau_2, 1)$, it is easy to think of combinations of T_{n_1}, T_{n_2} which remain on the same calibration scale. An effective combination is given in Sect. 19.1.2.

19.1.1 Desirable Properties of Statistical Evidence

Most statisticians, including us, would prefer an axiomatic approach to statistical evidence, but we provide an operational one. That is, guided by the above example, we state what properties we would like a measure of evidence to have, and then in specific problems show there are indeed statistics which come close to satisfying them. *The fact that it is an approximate theory in no way reduces its usefulness.* Normal approximations via the central limit theorem are ubiquitous in statistics, because they are useful in computing approximate p-values and confidence intervals. Similarly they are useful in providing evidence for alternative hypotheses.

Let θ represent an unknown real parameter for which it is desired to test $\theta = \theta_0$ against $\theta > \theta_0$, and let S_n be a test statistic based on n observations which rejects H_0 for large values of S_n. We want a measure of one-sided evidence $T_n = T_n(S_n)$ to satisfy:

E_1: The evidence T_n for a one-sided alternative is monotone increasing in S_n;
E_2: the distribution of T_n is normal for all values of the unknown parameters;
E_3: the variance $\mathrm{Var}[T_n] = 1$ for all values of the unknown parameters; and
E_4: the expected evidence $\tau(\theta) = \mathrm{E}_\theta[T_n]$ is increasing in θ from $\tau(\theta_0) = 0$.

In E_2, we require that the evidence *always* be unit normal, not only under the null hypothesis. As a consequence, the evidence proposed here carries much more information than results that are only true under the null hypothesis. For the prototypical model and $T_n(\bar{X}_n) = \sqrt{n}(\bar{X}_n - \theta_0)/\sigma_0$ all of the above properties hold exactly. Property E_1 is essential if the evidence is to remain a test statistic. In general, properties $E_2 - E_4$ will hold only approximately, but to a surprising degree, even for small sample sizes, provided one can find a *variance stabilizing transformation* (VST), of the test statistic S_n, $T_n = h_n(S_n) - h_n(\mathrm{E}_{\theta_0}[S_n])$ such that $\mathrm{Var}_\theta[T_n] \doteq 1$ for θ of interest. From now on, the symbol $\doteq$ signifies an approximate equality up to an error of smaller order in n. Since the variance of S_n is usually of order n^{-1}, the VST can usually be chosen as $h_n(\cdot) = \sqrt{n}h(\cdot)$.

(Kulinskaya et al. 2008, denoted KMS in the following) propose a measure of evidence in favor of alternative hypotheses that is based on a transformation of the usual test statistic to a normal translation family with unit variance, and provide numerous applications of it to standard problems of meta-analysis. Our purpose here is to explain in more detail why we advocate this particular definition. Connections with other measures of evidence, such as the p-value and Bayes factor, are given in Morgenthaler and Staudte (2012).

It turns out that the *expected KMS evidence*, when dealing with a sample of size n instead of a single observation, is equal to a product of two terms, the square root of n and a quantity $\mathcal{K}$ whose value indicates the difficulty in distinguishing the null density f_{θ_0} from an alternative density f_{θ_1}. This second term is the key to understanding and implementing inferential procedures (see Sect. 19.1.2 for details).

We restrict attention to a real-valued parametric family $f_\theta(x)$, where the testing problem of interest is $\theta = \theta_0$ against $\theta > \theta_0$. The elements of a traditional test are the test statistic S_n and its distribution under the null. To obtain a measure of evidence one needs a monotone transformation $T_n = h_n(S_n)$, which stabilizes the variance and is such that the distribution of T_n is approximately normal for all parameter values θ, *not only for the null value* θ_0. When the observation x is a realization of $X \sim f_{\theta_0}$, the likelihood ratio statistic on average favors f_{θ_0}, which means that $\mathrm{E}_{\theta_0}[\log(f_{\theta_0}(X)/f_\theta(X))] > 0$. This is a good measure of the difficulty in distinguishing f_{θ_0} from f_θ based on data from f_{θ_0}. It turns out that the symmetrized version of this quantity, the Kullback–Leibler Divergence, is closely linked to the function $h_n(\cdot)$.

Beginning with Fisher (1915), many statisticians have investigated “normalizing” a family of distributions through a transformation which often simultaneously stabilizes the variance, see the Wald Memorial Lecture by Efron (1982). As he points out, the purpose of transforming a test statistic so that its distribution is a normal translation family is both aesthetic (to gain insight) and practical (to easily obtain a confidence interval for an unknown parameter). To these desirable properties, we would add that this calibration scale is ideally suited for meta-analysis, because it allows for cancelation of evidence from conflicting studies, and facilitates combination of evidence obtained from several studies. Concerning this last point, the established theory of meta-analysis (see Becker 1997 or Thompson 1998), is a large-sample theory that is not very reliable for small sample sizes. Its implementation depends on estimators of weights and these estimators can be highly variable even for moderate sample sizes. By using variance stabilization first, researchers can apply the meta-analytic theory with much more confidence because, after transformation, no weights need to be estimated.

There is a constructive method for finding potential VSTs, see, for example, Bickel and Doksum (1977, p. 32) or Chap. 17 of Kulinskaya et al. (2008). These transformations are monotone increasing, so satisfy property E_1. They are defined only up to an additive constant, which may be chosen so that T satisfies property E_4. Variance stabilized statistics are often approximately normally distributed, and when they are so, the potential evidence T also “satisfies” E_2. The degree of satisfaction can be measured by simulation studies that show the VST leads to more accurate coverage of confidence intervals and more accurate estimates of power functions than the usual Central-Limit based approximations of the form

$(S_n - \mathrm{E}_{\theta_0}[S_n])/\sqrt{\mathrm{Var}_{\theta_0}[S_n]}$. Efron (1982) provides a constructive method for finding normalizing transformations.

19.1.2 Key Inferential Function

Suppose that one has in hand a measure of evidence T_n satisfying $E_1 - E_4$, at least asymptotically. In that case the expectation $\tau(\theta) = \mathrm{E}_\theta[T_n]$ summarizes the complete information. If we found T_n by application of a VST, that is, $T_n = h_n(S_n) - h_n(\mathrm{E}_{\theta_0}[S_n])$, then we can deduce $\tau(\theta) \doteq h_n(\mathrm{E}_\theta[S_n]) - h_n(\mathrm{E}_{\theta_0}[S_n])$, which can usually be written as $\tau(\theta) \doteq \sqrt{n}(h(\mathrm{E}_\theta[S_n]) - h(\mathrm{E}_{\theta_0}[S_n]))$.

Definition 19.1 Let T_n be a statistical evidence with $\tau(\theta) = \mathrm{E}_\theta[T_n] \doteq \sqrt{n}\mathcal{K}_{\theta_0}(\theta)$. Then $\mathcal{K}_{\theta_0}$ is called the *Key Inferential Function* or simply the *Key* for this statistical model and boundary value θ_0.

In the case of the normal shift model as given in the prototypical example, we found $\mathcal{K}_{\mu_0}(\mu) = (\mu - \mu_0)/\sigma_0$, which is often called the standardized effect and denoted by the symbol δ. In the case of a VST $h_n(\cdot) = \sqrt{n}h(\cdot)$, we have $\mathcal{K}_{\theta_0}(\theta) = h(\mathrm{E}_\theta[S_n]) - h(\mathrm{E}_{\theta_0}[S_n])$. This last expression is simply a centered version of the VST, where the centering assures the equality $\mathcal{K}_{\theta_0}(\theta_0) = 0$.

The *Key* contains all the essential information, and knowing it enables one to solve many routine statistical problems, such as

K_1: *Choosing sample sizes:* For testing $\theta = \theta_0$ against $\theta > \theta_0$ using a sample of n observations the expected evidence is $\tau(\theta) = \sqrt{n}\mathcal{K}_{\theta_0}(\theta)$ for each θ. To attain a desired expected evidence τ_1 against alternative θ_1 one can choose n_1 to be the smallest integer greater than or equal to $[\tau_1/\mathcal{K}_{\theta_0}(\theta_1)]^2$.

For the prototypical model, this means $n_1 \geq \{\tau_1/\delta_1\}^2$, where $\delta_1 = (\mu_1 - \mu_0)/\sigma_0$. Also, for this model the test statistic is $T_n = \sqrt{n}(\bar{X}_n - \mu_0)/\sigma_0 \sim N(\tau_1, 1)$, where $\tau_1 = \sqrt{n}\delta_1$. Hence the power $1 - \beta(\mu_1)$ of the level α test for μ_1 is exactly $1 - \beta(\mu_1) = P_{\mu_1}(T_n \geq z_{1-\alpha}) = \Phi(\tau_1 - z_{1-\alpha})$; that is, $\tau_1 = z_{1-\alpha} + z_{1-\beta(\mu_1)}$. Now substituting this expression for τ_1 into the lower bound for n_1 gives the well known expression $n_1 \geq \{\tau_1/\delta_1\}^2 = \sigma_0^2\{z_{1-\alpha} + z_{1-\beta(\mu_1)}\}^2/(\mu_1 - \mu_0)^2$.

K_2: *Power calculations:* A Neyman–Pearson level α test based on T_n has power $1 - \beta(\theta)$ against alternative θ given by

$$1 - \beta(\theta) \doteq \Phi\left(\sqrt{n}\mathcal{K}_{\theta_0}(\theta) - z_{1-\alpha}\right) \quad \text{or} \tag{19.1}$$

$$\sqrt{n}\mathcal{K}_{\theta_0}(\theta) = z_{1-\alpha} + z_{1-\beta(\theta)}. \tag{19.2}$$

Formula (19.1) often leads to more accurate power approximations than standard asymptotics, see Kulinskaya et al. (2008, Chap. 22). It follows that accurate choice of sample size to obtain power at a given level is possible. Formula (19.2) shows that the VST expected evidence is more basic than level and power: it can be partitioned into the sum of the probits of the false positive and false negative error rates.

K_3: *Confidence intervals:* A $100(1-\alpha)$ % confidence interval for θ is given by

$$\left[\mathcal{K}_{\theta_0}^{-1}\left(\frac{T_n - z_{1-\alpha/2}}{\sqrt{n}}\right), \mathcal{K}_{\theta_0}^{-1}\left(\frac{T_n + z_{1-\alpha/2}}{\sqrt{n}}\right)\right], \tag{19.3}$$

where $\mathcal{K}^{-1}$ is the inverse function to $\mathcal{K}$.

For the prototypical model the *Key* is $\mathcal{K}_{\mu_0}(\mu) = (\mu - \mu_0)/\sigma_0 = \delta$, so $\mathcal{K}_{\mu_0}^{-1}(\kappa) = \sigma_0\kappa + \mu_0$. Substituting $T_n = \sqrt{n}(\bar{X}_n - \mu_0)/\sigma_0$ into (19.3) produces the confidence interval $[\bar{X}_n - z_{1-\alpha/2}\sigma_0/\sqrt{n}, \bar{X}_n + z_{1-\alpha/2}\sigma_0/\sqrt{n}]$.

K_4: *Meta-analysis for the fixed effects model:* Given independent $T_1, \ldots, T_K$, where $T_k = T_{n_k} \sim N(\tau_k, 1)$ with $\tau_k = \sqrt{n_k}\mathcal{K}_{\theta_0}(\theta)$, each evidence for $\theta > \theta_0$, let

$$T_{1:K} = \frac{\sqrt{n_1}T_1 + \cdots + \sqrt{n_K}T_K}{\sqrt{N_K}}, \tag{19.4}$$

where $N_K = \sum_k n_k$. Then $T_{1:K} \sim N(\tau_{1:K}, 1)$, with $\tau_{1:K} = \sqrt{N_K}\mathcal{K}_{\theta_0}(\theta)$, is the combined evidence for $\theta > \theta_0$, and a $100(1-\alpha)$ % confidence interval for θ based on all the evidence is found by replacing the T_n of (19.3) by $T_{1:K}$.

For the prototypical model, where $T_k = \sqrt{n_k}(\bar{X}_k - \mu_0)/\sigma_0 \sim N(\tau_k, 1)$ with $\tau_k = \sqrt{n_k}\mathcal{K}_{\mu_0}(\mu)$ and $\mathcal{K}_{\mu_0}(\mu) = (\mu - \mu_0)/\sigma_0$, one has $T_{1:K} = \sqrt{N_K}(\bar{\bar{X}} - \mu_0)/\sigma_0$. Here, $\bar{\bar{X}}$ is the mean of all $N_K = \sum_k n_k$ observations.

Note that if the initial statistical model is reparameterised in terms of $\eta = \eta(\theta)$, where $\eta(\cdot)$ is a strictly increasing function, then the *Key* $\mathcal{K}_{\eta_0}(\eta)$ becomes the composition of $\mathcal{K}_{\theta_0}(\theta)$ with the inverse reparametrization $\theta = \theta(\eta)$, that is, $\mathcal{K}_{\eta_0}(\eta) = \mathcal{K}_{\theta(\eta_0)}(\theta(\eta))$. The transformation to the “right parameter” $\eta = \mathcal{K}_{\theta_0}(\theta)$, for example, leads to $\mathcal{K}_{\eta_0}(\eta) = \eta$, where $\eta_0 = 0 = \mathcal{K}_{\theta_0}(\theta_0)$.

For all the above reasons, the *Key* appears to contain all the information required for inference in one-parameter families, and this claim is supported by the material in the next Sect. 19.2. In it, we describe the very strong link between the *Key* and the Kullback–Leibler Divergence for exponential families. In Sect. 19.3, we illustrate many of the above results for the non-central chi-squared family, which is not an exponential family. In Sect. 19.4, we summarize the results and describe areas for future research.

19.2 Connection to the Kullback–Leibler Divergence

Kullback (1968) is a well-written and highly informative book whose principal topic is the following measure of information

$$I(\theta_0 : \theta_1) = \mathbf{E}\left[\log\left(\frac{f_{\theta_0}(X)}{f_{\theta_1}(X)}\right)\right], \quad \text{where } X \sim f_{\theta_0}.$$

This quantity is the average value of the log likelihood ratio when choosing between the model densities f_{θ_0} and f_{θ_1} with data X that is generated by f_{θ_0}. The logarithm of the likelihood ratio $\log(f_{\theta_0}(x)/f_{\theta_1}(x))$ is taken as the information in an observation $X = x$ for discrimination in favor of $X \sim f_{\theta_0}$ against $X \sim f_{\theta_1}$ (Kullback 1968, p. 5). A variety of strong arguments give backing to this choice.

Definition 19.2 The symmetrized information, defined as $J(\theta_0, \theta_1) = I(\theta_0 : \theta_1) + I(\theta_1 : \theta_0)$ is called the *Kullback–Leibler Divergence (KLD)* (see Kullback 1968, p. 6).

Kullback's terminology has been modified over the years, and now $I(\theta_0 : \theta_1)$ is often called the *divergence* or *directed divergence* and $J(\theta_0, \theta_1)$ the *symmetrized divergence*. When the likelihood ratio test is performed with n independent observations, both I and J for discriminating will be multiplied by n. Thus in most of the examples and theory to follow, we can omit the sample size.

19.2.1 Example 1: Normal Model

We begin with a return to the prototypical model in which there are no surprises, but the generality soon becomes clear. If f_{μ_0} and f_{μ_1} are normal densities with equal variances σ_0^2, but unequal means μ_0 and μ_1, the Kullback–Leibler Information is

$$I(\mu_0 : \mu_1) = \mathrm{E}\left[\frac{1}{2}\left(\frac{(X-\mu_1)^2}{\sigma_0^2} - \frac{(X-\mu_0)^2}{\sigma_0^2}\right)\right], \quad \text{where } X \sim f_0.$$

Therefore, $I(\mu_0 : \mu_1) = \frac{1}{2}(1 + (\mu_1 - \mu_0)^2/\sigma_0^2 - 1)$ and $J(\mu_0, \mu_1) = (\mu_1 - \mu_0)^2/\sigma_0^2 = \delta^2$. The Kullback–Leibler Divergence is equal to the square of the standardized effect δ. The information for discrimination is thus equal to the square of the Key Inferential Function for the z test of the null hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$, namely $\mathcal{K}_{\mu_0}(\mu_1) = \delta$, found in Sect. 19.1.

The above example can be extended to the case of evidence for alternative $\theta > \theta_0$ to the null $\theta = \theta_0$, for which the *Key* is $\mathcal{K}_{\theta_0}(\theta)$, where we now drop the subscript on the parameter in the alternative $\theta > \theta_0$. We also write $J_{\theta_0}(\theta)$ for $J(\theta_0, \theta)$ to emphasize that θ_0 is fixed and θ is any value in the alternative $\theta > \theta_0$. The Kullback–Leibler Divergence (KLD) between the models $\mathbf{N}(0, 1)$ and $\mathbf{N}(\mathcal{K}_{\theta_0}(\theta), 1)$ is by the previous example equal to $\mathcal{K}^2_{\theta_0}(\theta)$. This suggests that we can find the *Key* and the VST $h(\cdot)$ by computing the KLD, because

$$\mathcal{K}_{\theta_0}(\theta) \approx \sqrt{J_{\theta_0}(\theta)}\,\mathrm{sgn}(\theta - \theta_0). \tag{19.5}$$

Common examples for which this approximation is excellent for θ in a large neighborhood of the null value θ_0 are the Poisson, exponential, binomial, and the correlation coefficient of bivariate normal. It is also true for the non-central t, see Morgenthaler and Staudte (2012), and the non-central chi-square models, see Sect. 19.3.

19.2.2 Result for Exponential Families

Let X have density of the form $f(x \mid \eta) = g(x)\exp\{\eta x - k(\eta)\}$ for x in an interval not depending on η. These densities for X are called an exponential family with natural parameter η; see Severini (2000) for background material. We further assume that $\mathrm{Var}_\eta[X] > 0$ for all η. We want to compare the Kullback–Leibler Symmetrized Divergence with the square of the Key Inferential Function for this class of models. As a Corollary, we will compare the *Key* itself with the signed square root of the divergence.

The derivatives of the function k give the cumulants of X; so that

$$\begin{aligned}\mu &= \mathrm{E}_\eta[X] = \kappa_1(\eta) = k'(\eta)\\ \sigma^2 &= \mathrm{Var}_\eta[X] = \kappa_2(\eta) = k''(\eta)\\ \mathrm{E}_\eta\big[(X-\mu)^3\big] &= \kappa_3(\eta) = k'''(\eta).\end{aligned} \tag{19.6}$$

Now $\mu = k'(\eta)$ has positive derivative, and therefore a monotone increasing inverse $\eta = (k')^{-1}(\mu)$, so all the cumulants of X can be written as functions of μ. For example, $\sigma^2(\mu) = k'' \circ (k')^{-1}(\mu)$.

The Kullback–Leibler Information about $f(\cdot \mid \eta)$ when $f(\cdot \mid \eta_0)$ is the density of X is

$$I(\eta_0 : \eta) = \mathrm{E}_{\eta_0}\big[\ln\big(f\big(X \mid \eta_0\big)/f\big(X \mid \eta\big)\big)\big] = (\eta_0 - \eta)k'(\eta_0) - k(\eta_0) + k(\eta).$$

Therefore, the Divergence is

$$\begin{aligned}J_{\eta_0}(\eta) &= (\eta - \eta_0)\big\{k'(\eta) - k'(\eta_0)\big\}\\ &= \big\{\big(k'\big)^{-1}(\mu) - \big(k'\big)^{-1}(\mu_0)\big\}(\mu - \mu_0) = J_{\mu_0}(\mu).\end{aligned}$$

If a VST $h(X)$ for X exists which has variance $\mathrm{Var}[h(X)] \doteq 1$, it must satisfy $h'(\mu) = 1/\sigma(\mu)$, and the *Key* for testing $\mu = \mu_0$ against $\mu > \mu_0$ is defined by $\mathcal{K}_{\mu_0}(\mu) = h(\mu) - h(\mu_0)$.

Proposition 19.1 *Suppose the model is a one-parameter exponential family and let $J_{\mu_0}(\mu)$ denote the Kullback–Leibler Divergence, whereas $\mathcal{K}_{\mu_0}(\mu)$ is the Key Inferential Function. It follows that*

$$J_{\mu_0}(\mu) = \mathcal{K}^2_{\mu_0}(\mu)\big\{1 + C_2(\mu - \mu_0)^2/2! + O\big(|\mu - \mu_0|^3\big)\big\}$$

and

$$\mathrm{sign}(\mu - \mu_0)\sqrt{J_{\mu_0}(\mu)} = \mathcal{K}_{\mu_0}(\mu)\left\{1 + \frac{1}{2}C_2(\mu - \mu_0)^2/2! + O\big(|\mu - \mu_0|^3\big)\right\},$$

where $C_2 = \kappa_3^2(\mu_0)/\{24\sigma^8(\mu_0)\}$.

A proof is given in Morgenthaler and Staudte (2012).

For contiguous alternatives $\theta_n = \theta_0 + O(1/\sqrt{n})$, the relative error in the approximation is of order $O(1/n)$. Thus, the approximation remains useful for alternatives that are much further removed from the null value than the contiguous ones.

The procedure based on variance stabilization is applicable beyond the context of exponential families. The basic idea of approximating a test problem by a normal translation family is not new and it is well known that many hypothesis testing procedures, which reject for large values of S_n, take this form for large sample sizes n and contiguous alternatives. This is true in the sense that the power of the level α test of $\theta = \theta_0$ against the alternatives $\theta > \theta_0$ is approximately equal to $\Phi(z_\alpha + \sqrt{n}e(\theta_0)(\theta - \theta_0))$, where z_α denotes the α quantile of the standard normal distribution and $\sqrt{n}e(\theta_0) = \mu'(\theta_0)/\sigma(\theta_0) > 0$ describes the efficacy of the test statistic, where $\mu(\theta)$ and $\sigma^2(\theta)$ are the mean and variance of S_n. For the variance stabilized test statistic $T_n = h_n(S_n)$, the simpler formula $\Phi(z_\alpha + \sqrt{n}\mathcal{K}_{\theta_0}(\theta))$ is obtained and as we have seen, this gives a good approximation beyond contiguous alternatives. In order that these two formulae agree in a neighborhood of θ_0, it must be true that $\frac{d}{d\theta}\mathcal{K}_{\theta_0}(\theta)$, evaluated at the null value θ_0, is equal to $\mu'(\theta_0)/\sigma(\theta_0)$. Because the *VST* satisfies $\frac{d}{d\mu}h(\theta_0) = 1/\sigma(\theta_0)$, this is indeed the case.

19.2.3 Example 2: Poisson Model

Let $X \sim \text{Poisson}(\lambda)$ and find the evidence for $\lambda > \lambda_0$ when the null hypothesis is $\lambda \leq \lambda_0$. An elementary calculation gives $I(\lambda_0 : \lambda) = \lambda - \lambda_0 + \lambda_0 \log(\lambda_0/\lambda)$, which implies that $J_{\lambda_0}(\lambda) = (\lambda - \lambda_0)\log(\lambda/\lambda_0)$. The classical VST for the Poisson model leads to $\mathcal{K}_{\lambda_0}(\lambda) = \sqrt{4\lambda} - \sqrt{4\lambda_0}$. The graphs of $\mathcal{K}_{\lambda_0}(\lambda)$ and $\sqrt{J_{\lambda_0}(\lambda)}\,\text{sgn}(\lambda - \lambda_0)$ are in agreement in a relatively large neighborhood of λ_0, regardless of its value. To check this, consider the parametrization $\lambda = \lambda_0 + (\lambda - \lambda_0) = \lambda_0 + \Delta$ for which we have $J_{\lambda_0}(\lambda) = \Delta\log(1 + \Delta/\lambda_0) = (\Delta^2/\lambda_0)(1 - \Delta/(2\lambda_0))$, while $\mathcal{K}_{\lambda_0}(\lambda) = 2(\sqrt{\lambda_0 + \Delta} - \sqrt{\lambda_0}) = \Delta/\sqrt{\lambda_0} - \Delta^2/(4\lambda_0^{3/2})$. The leading term of the signed root of J and of the *Key* is $\Delta/\sqrt{\lambda_0}$, which is the standardization obtained by dividing the raw effect $\lambda - \lambda_0$ by the standard error at the null hypothesis. We leave it to the reader to check that the next order term also is in agreement. The classical VST suggests that when $\lambda_0 = 1$, the correct parameter to use for testing and evaluating evidence is $\eta = 2(\sqrt{\lambda} - 1)$, while the KLD gives $\sqrt{(\lambda - 1)\log(\lambda)}\,\text{sgn}(\lambda - 1)$.

19.3 Non-central Chi-Squared Family

In this section, we illustrate some of the results from Sects. 19.1 and 19.2 in the context of the chi-squared family with known degrees of freedom and unknown non-centrality parameter. This model is not an exponential family.

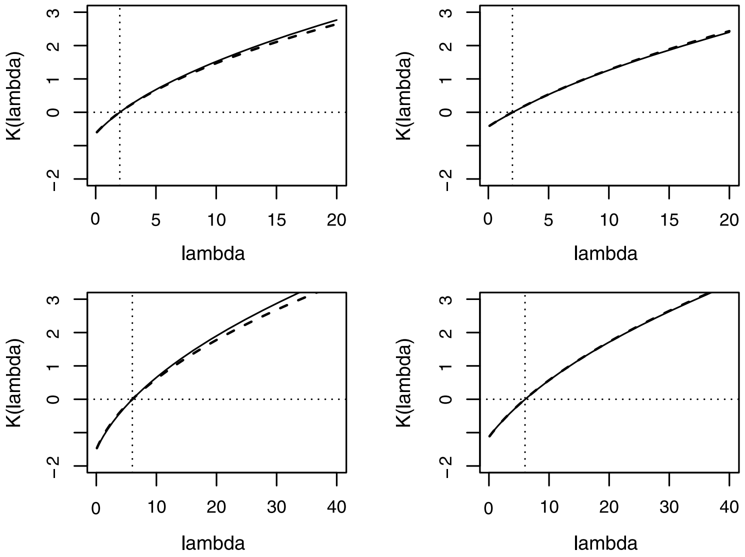


Fig. 19.1 In all plots *the solid line* depicts the graph of $\mathcal{K}_{\lambda_0}(\lambda)$, for the $\chi^2_\nu(\lambda)$ model, when testing $\lambda \le \lambda_0$ against $\lambda > \lambda_0$. *The dashed line* that approximates it is the signed square root of the Kullback–Leibler symmetrized divergence $\sqrt{J_{\lambda_0}(\lambda)}\,\mathrm{sgn}(\lambda-\lambda_0)$. The latter is computed by Monte Carlo integration on R. For *the two left hand plots* $\nu = 3$ and in *the upper plot* $\lambda_0 = 2$ while in *the bottom plot* $\lambda_0 = 6$. For the *two right hand plots* $\nu = 9$ and in *the upper plot* $\lambda_0 = 2$ while in *the bottom plot* $\lambda_0 = 6$. *The dotted vertical lines* mark the null hypothesis

19.3.1 Comparing the KLD with the Key

Let $X \sim \chi^2_\nu(\lambda)$ have the non-central chi-squared distribution with ν degrees of freedom and non-centrality parameter λ. In most applications, ν is known and λ is unknown. It is not possible to compute the Kullback–Leibler symmetrized divergence (KLD) between $\chi^2_\nu(\lambda_0)$ and $\chi^2_\nu(\lambda_1)$ analytically, but because of the well-known VST, we think that it has to be

$$J(\lambda_0;\lambda_1) \doteq (\sqrt{\lambda_1+\nu/2}-\sqrt{\lambda_0+\nu/2})^2. \qquad (19.7)$$

The approximation (19.7) is confirmed by computational results for many choices of ν, λ_0 and λ_1, some of which are presented in Fig. 19.1. But first we show the motivation for the conjecture by finding the *Key* for the evidence in X when testing $\lambda \le \lambda_0$ against $\lambda > \lambda_0$. Using the fact that $\mathrm{E}[X] = \nu+\lambda$ and $\mathrm{Var}[X] = 2\nu+4\lambda$, one can write $\mathrm{Var}[X] = g(\mathrm{E}[X])$, where $g(t) = 4t-2\nu$. Its inverse square root has indefinite integral

$$h-\nu(x) = \int^x \frac{dt}{\sqrt{4t-2\nu}} = \sqrt{x-\nu/2}+c. \qquad (19.8)$$

Thus by the standard method (Bickel and Doksum 1977, p. 32), $h_\nu(x)$ is a potential VST for X. It is only defined for $x > \nu/2$, but this is not a practical restriction because

$$P_{\nu,\lambda}(X \le \nu/2) \le P_{\nu,0}(X \le \nu/2) \approx \Phi\left(-\frac{\sqrt{\nu}}{2\sqrt{2}}\right), \tag{19.9}$$

which is negligible even for moderate ν. The approximation of $\mathrm{E}[h(X)]$ by $\sqrt{\mathrm{E}[X] - \nu/2} + c$ leads to (19.7).

19.3.2 Tests for the Non-centrality Parameter

Given $X_1, \ldots, X_n$ i.i.d. with $X_i \sim \chi^2_\nu(\lambda)$, it is desired to test the null $\lambda = \lambda_0$ against $\lambda > \lambda_0$ using as test statistic the sample mean $\bar{X}_n$. Any VST is derived as above to be $h_n(\bar{X}_n) = \sqrt{n}\sqrt{\bar{X}_n - \nu/2} + c$. To convert this into evidence T_n for $\lambda > \lambda_0$ we need to choose c so that $\mathrm{E}[T_n] = \mathrm{E}[h_n(\bar{X}_n)]$ is monotone increasing in λ with value 0 at the boundary $\lambda = \lambda_0$. To a first approximation, $\mathrm{E}[\sqrt{\bar{X}_n - \nu/2}] = \sqrt{\lambda + \nu/2}$ so we choose $c = -\sqrt{n}\sqrt{\lambda_0 + \nu/2}$. Then

$$\mathrm{E}[T_n] \doteq \sqrt{n}[\sqrt{\lambda + \nu/2} - \sqrt{\lambda_0 + \nu/2}]. \tag{19.10}$$

It remains to check that T_n is approximately normal with variance near 1 and this is left to the reader. Other important results are that the evidence grows with the square root of the sample size and the *Key* function is monotone increasing in λ from 0 at the null. The *Key* function evidently is $\mathcal{K}_{\lambda_0}(\lambda) = \sqrt{\lambda + \nu/2} - \sqrt{\lambda_0 + \nu/2}$. Now it is apparent, in view of Proposition 19.1, how the conjecture (19.7) arises, even though the non-central chi-squared distribution is not an exponential family.

Figure 19.1 shows some examples of the approximation (19.7). Even for $\nu = 3$ (left-hand plots) the approximation is good near the null; and the approximations appear to improve with ν. This means that we can use the simple expression $\mathcal{K}_{\lambda_0}(\lambda) = \sqrt{\lambda + \nu/2} - \sqrt{\lambda_0 + \nu/2}$ for the *Key* to carry out inference for λ as described in Sect. 19.1. Further, we know that the *Key* is a good approximation to the signed square root of the KLD between null and alternative hypothesized distributions, at least for a large neighborhood of λ_0.

While the above ideas are straightforward, we do not always have n independent observations on a chi-squared family; rather the non-central chi-squared distribution arises through a consideration of K groups, as described in the next subsection.

19.3.3 Between Group Sum of Squares (for Known Variance)

For each group $k = 1, \ldots, K$ let $\mathbf{X}'_k = [X_{k1}, X_{k2}, \ldots, X_{k,n_k}]$ denote a sample of n_k observations, each with distribution $N(\mu_k, 1)$. Also assume the elements of $\mathbf{X}' = [\mathbf{X}_1, \ldots, \mathbf{X}_K]$ are independent. Further introduce the total sample size $N = \sum_k n_k$, the sample proportions $q_k = n_k/N$, the kth sample mean $\bar{X}_k$, the overall sample mean $\bar{X} = \sum_k q_k \bar{X}_k$, its expectation $\mu = \sum_k q_k \mu_k$ and the parameter $\lambda = N \sum_k q_k (\mu_k - \mu)^2$. Then the between group sum of squares $Y = N \sum_k q_k (\bar{X}_k - \bar{X})^2 \sim \chi^2_\nu(\lambda)$, where $\nu = K - 1$, see Kulinskaya et al. (2008, Sect. 22.1). The ratio $\theta = \lambda/N = \sum_k q_k(\mu_k - \mu)^2$ depends only on the *relative* sample sizes q_k, and measures the variability of the group means μ_k using a weighted sum of squared deviations from the weighted mean μ, with weights q_k.

Let the test statistic be $S = Y/N$. The transformation to evidence for $\theta > \theta_0$ is then $T = \sqrt{N}[\sqrt{S - \nu/(2N)} - \sqrt{\theta_0 + \nu/(2N)}]$. Further, introduce the parameter $r = \nu/N = (K-1)/N$; the mean and variance of S in this notation are $\mathrm{E}[S] = \theta + r$ and $\mathrm{Var}[S] = (4\theta + 2r)/N$. The expected evidence for $\theta \geq \theta_0$ become $\mathrm{E}_\theta[T] \doteq \sqrt{N}\mathcal{K}_{\theta_0,N}(\theta)$, with the *Key* given by

$$\mathcal{K}_{\theta_0,N}(\theta) \doteq \sqrt{\theta + r/2} - \sqrt{\theta_0 + r/2} - \frac{1}{2N\sqrt{\theta + r/2}}. \tag{19.11}$$

This shows that the expected evidence is monotone increasing in θ for $\theta > \theta_0$, and is approximately 0 at $\theta = \theta_0$. For fixed θ it grows with $\sqrt{N}$. Also, for fixed θ, if $r = (K-1)/N$ remains fixed with increasing N, the correction term becomes negligible and the *Key* is essentially the first two terms of (19.11). If $K = o(N)$ as $N \to \infty$, then $r \to 0$ and the Key approaches $\mathcal{K}_{\theta_0,+\infty}(\theta) = \sqrt{\theta} - \sqrt{\theta_0}$.

Confidence Intervals for the Non-centrality Parameter To obtain the confidence bounds of (19.3), we need to solve for $\theta = \mathcal{K}^{-1}_{\theta_0,N}(u)$ Setting $c = -\sqrt{\theta_0 + r/2}$ we start with

$$u = \mathcal{K}_{\theta_0,N}(\theta) = \sqrt{\theta + r/2} + c - \frac{1}{2N\sqrt{\theta + r/2}}. \tag{19.12}$$

Solving this quadratic in θ yields

$$\mathcal{K}^{-1}_{\theta_0,N}(u) = \frac{1}{2}\left[\frac{1}{N} + \left\{(u-c)^2 - r\right\} + \left\{(u-c)^4 + \frac{2(u-c)^2}{N}\right\}^{1/2}\right]. \tag{19.13}$$

Evaluating this function at $u_\pm = (T \pm z_{0.975})/\sqrt{N}$, for $T = \sqrt{N}[\sqrt{S - \nu/(2N)} + c]$ yields the 95 % confidence interval for θ in terms of the test statistic $S = Y/N = \sum_k q_k(\bar{X}_k - \bar{X})^2$. For convenience, we note that $u_\pm - c = \sqrt{S - r/2} \pm z_{0.975}/\sqrt{N}$.

The performance of T and confidence intervals for θ based on it were examined by generating 100,000 simulations of $Y = NS \sim \chi^2_\nu(\lambda)$ for various choices of $\nu = K - 1$ and N, and then computing the average bias $T - \sqrt{N}\mathcal{K}_{\theta_0,N}(\theta)$ (which is free of θ_0), the average standard deviation SD[T], the one-sided 95 % confidence bound empirical coverage, and finally the two-sided 95 % confidence interval empirical coverage probabilities. These results are plotted as a function of θ over the range [0, 3] in Fig. 19.2. In the above derivation of confidence intervals, we included a bias term in the *Key* to see if the resulting confidence intervals had better coverage than when we used the simpler the simpler *Key* $\mathcal{K}_{\theta_0}(\theta) = \sqrt{\theta + r/2} - \sqrt{\theta_0 + r/2}$. However, one only loses a little in accuracy of coverage probabilities and the derivation of the confidence interval is much quicker by the standard method K3 of Sect. 19.1.2.

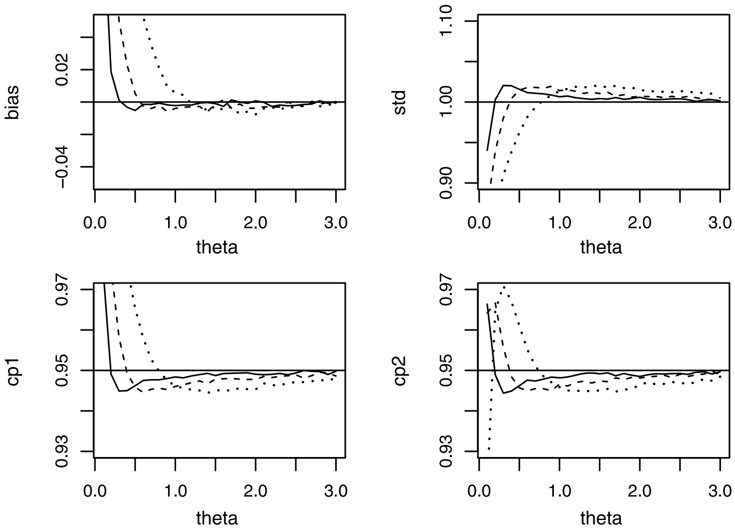


Fig. 19.2 In the first row of plots above are shown the empirical biases and standard deviation of T for $\nu = K - 1 = 4$ degrees of freedom in Example 2 of Sect. 19.3. The results correspond to $N = 10$ (*dotted line*), $N = 20$ (*dashed line*), and $N = 40$ (*solid line*). *The second row* of plots gives the empirical coverage probabilities of nominal 95 % upper confidence bounds and 95 % confidence intervals

19.4 Conclusions

We have shown that it is often possible and practical to define an evidence T in favor of alternatives. This statistic is based on the idea of variance stabilization and the mean function of this evidence is closely related to the Kullback–Leibler divergence (KLD). Investigating the generality of this result merits further research.

In general, it may be said that the KLD gives insights into a variety of inferential questions and deserves renewed attention by statisticians. In the following, we give two other examples that show the power of the KLD in revealing underlying structure. When the densities to be compared are $f_i(x) = f(x/\sigma_i)/\sigma_i$, one has $\text{KLD}(\sigma_1, \sigma_2) = \text{KLD}(1, \sigma_2/\sigma_1)$—the ratio of the scales is the essential parameter. To be more precise, we have to compute the KLD. If the underlying density is normal, one obtains $\text{KLD}(\sigma_1, \sigma_2) = \frac{1}{2}(\sigma_2/\sigma_1 - \sigma_1/\sigma_2)^2$. Reparametrizing to $\sigma_2 = (1 + \Delta)\sigma_1$, we have $\frac{1}{2}(1 + \Delta - 1/(1 + \Delta))^2$ for the value of the KLD. This expands into $\frac{1}{2}(1 + \Delta - [1 - \Delta + \Delta^2 + O(\Delta^3)])^2 = \frac{1}{2}(2\Delta - \Delta^2 + O(\Delta^3))^2$. The square root for $\Delta > 0$ leads to the *Key* $\sqrt{2}\Delta - \Delta^2/\sqrt{2} + O(\Delta^3)$, which is up to this order the same as $\sqrt{2}\log(1 + \Delta) = \sqrt{2}(\log(\sigma_2) - \log(\sigma_1))$. Thus, the transformed parameter obtained through the signed root of the KLD is simply the logarithm and furthermore, the test statistic is based on the difference. This is, of course, simply related to the fact that if the observed random variable $Y = \sigma X_0$, then $\log(Y) = \log(X_0) + \log(\sigma)$, which transforms the model into location-form.

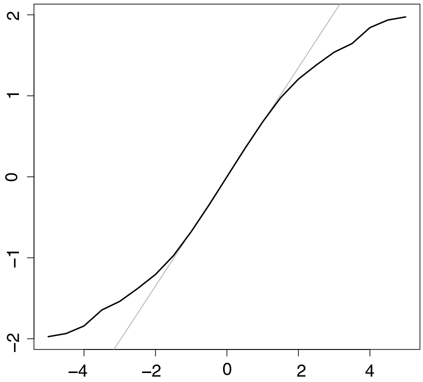


Fig. 19.3 *The dark curve* shows the signed root of the KLD for two standard Cauchy densities with a translational shift between them. The values were computed by Monte Carlo simulation. The value of the shift is indicated on the x-axis. *The grey line* has a slope equal to the ratio of the normal upper quartile divided by the Cauchy upper quartile, which can serve as an estimator of the scale change when switching the standard normal to the standard Cauchy density. For small shifts, there is but a tiny difference between the straight line and the root of the Cauchy KLD. For large shifts, the Cauchy KLD grows at a slower pace and turns out to be sub-linear

Another example concerns robust, heavy-tailed models. When comparing $f_i(x) = f((x - \mu_i)/\sigma_0)/\sigma_0$, it is easy to show that the KLD only depends on $\delta = (\mu_2 - \mu_1)/\sigma$. As we have seen in our prototypical example, the KLD has value δ^2 for the normal shift model. What happens, if one moves to a heavy-tailed density? Figure 19.3 shows the case of the Cauchy density. It turns out that the amount of information available for small δ remains linear in δ and a loss of information only occurs for large values. Thus, with appropriate estimators of δ, no loss of information due to heavy-tails occurs. The loss is only due to the difficulty in estimating δ. As robust theory shows, it is possible to construct compromise estimators that exploit this underlying information successfully for a wide range of tail behaviors. A similar loss of information for large values of δ occurs in the central Student-t model with unknown scale σ and a smallish number of degrees of freedom.

Even though we have only considered cases where the underlying parameter takes real values, extensions to multidimensional parameters are possible and this problem is open to further investigation. It would also be of interest to consider examples where the evidence is multidimensional.

References

Becker, B. J. (1997). P-values, combination of. In S. Kotz (Ed.), *Encyclopedia of statistical sciences* (Vol. 1, pp. 448–453). New York: Wiley.

Bickel, P. J., & Doksum, K. A. (1977). *Mathematical statistics: basic ideas and selected topics*. Oakland: Holden-Day.

Efron, B. (1982). Transformation theory: how normal is a family of distributions? *The Annals of Statistics*, *10*, 323–339.

Evans, M. (2000). Comment: on the probability of observing misleading statistical evidence. *Journal of the American Statistical Association*, *95*, 768–769.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*, 507–521.

Kulinskaya, E., Morgenthaler, S., & Staudte, R. G. (2008). *Meta analysis: a guide to calibrating and combining statistical evidence*. Sussex: Wiley.

Kullback, S. (1968). *Information theory and statistics*. New York: Dover.

Malzahn, U., Bohning, D., & Holling, H. (2000). Nonparametric estimation of heterogeneity variance for the standardized difference used in meta-analysis. *Biometrika*, *87*, 619–632.

Morgenthaler, S., & Staudte, R. G. (2012). Advantages of variance stabilization. *Scandinavian Journal of Statistics.*, *39*(4), 714–728. doi:10.1111/j.1467-9469.2011.00768.x.

Royall, R. (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association*, *95*, 760–768.

Severini, T. A. (2000). *Likelihood methods in statistics*. London: Oxford University Press.

Thompson, S. G. (1998). Meta analysis of clinical trials. In P. Armitage & T. Colton (Eds.), *Encyclopedia of biotatistics* (pp. 2570–2579). London: Wiley.

Chapter 20
Concepts and a Case Study for a Flexible Class of Graphical Markov Models

Nanny Wermuth and David R. Cox

20.1 Introduction

To observe and understand relations among several features of individuals or objects is one of the central tasks in many substantive fields of research, including the medical, social, environmental and technological sciences. Statistical models can help considerably with such tasks provided they are both flexible enough to apply to a wide variety of different types of situation and precise enough to guide us in thinking about possible alternative relationships. This requires in particular joint responses, which contain continuous random variables, discrete random variables or both types, in addition to only single responses.

Causal inquiries, the search for causes and their likely consequences, motivate much empirical research. They rely on appropriate representations of relevant pathways of dependence as they develop over time, often called data generating processes. Causes which start pathways with adverse consequences may be called risk factors or risks. Knowing relevant pathways offers in principle the opportunity to intervene, aiming to stop the accumulation of some of the risks, and thereby to prevent or at least alleviate their negative consequences.

Properties of persons or objects and features, such as attitudes or behavior of individuals, which can vary for the units or individuals under study, form the variables that are represented in statistical models. A relationship is called a strong positive dependence if knowing one feature makes it much more likely that the other feature

N. Wermuth (✉)
Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden
e-mail: wermuth@chalmers.se

N. Wermuth
International Agency of Research on Cancer, Lyon, France

D.R. Cox
Nuffield College, Oxford, UK
e-mail: david.cox@nuffield.ox.ac.uk

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_20, © Springer-Verlag Berlin Heidelberg 2013

is present as well. If, however, prediction of a feature cannot be improved by knowing the other, then the relation of the two is called an independence. Whenever such relations only hold under certain conditions, then they are qualified to be conditional dependences or independences.

Graphs, with nodes representing variables and edges indicating dependences, serve several purposes. These include to incorporate available knowledge at the planning stage of an empirical study, to summarize aspects important for interpretation after detailed statistical analyses and to predict, when possible, effects of interventions, of alternative analyses of a given set of data or of changes compared to results from other studies with an identical core set of variables.

Corresponding statistical models are called graphical Markov models. Their graphs are simple when they have at most one edge for any variable pair even though there may be different types of edge. The graphs can represent different aspects of pathways, such as the conditional independence structure, the set of all independence statements implied by a graph, or they indicate which variables are needed to generate joint distributions. In the latter case, the graph represents a research hypothesis on variables that make an important contribution. Theoretical and computational work has progressed strongly during the last few years.

In the following, we give first some preliminary considerations. Then we describe some of the history of graphical Markov models and the main features of their most flexible subclass, called traceable regressions. We illustrate some of the insights to be gained with sequences of joint regressions, that turn out to be traceable in a prospective study of child development, now known as the Mannheim Study of Children at Risk.

20.2 Several Preliminary Considerations

Graphical Markov models are of interest in different contexts. In the present paper, we stress data analysis and interpretation. From this perspective, a number of considerations arise. In a given study, we have objects or individuals, here children, and their appropriate selection into the study is important. Each individual has properties or features, represented as variables in statistical models.

A first important consideration is that for any two variables, either one is a possible outcome to the other, regarded as possibly explanatory, or the two variables are to be treated as of equal standing. Usually, an outcome or response refers to a later time period than a possibly explanatory feature. In contrast, an equal standing of two or more features is appropriate when they refer to the same time period or all of them are likely to be simultaneously affected by an intervention.

On the basis of this, we typically organize the variables for planned statistical analyses into a series of blocks, often corresponding to a time ordering. All relations between variables within a same block are undirected, whereas those between variables in different blocks are directed in the way described.

An edge between two nodes in the graph, representing a statistical dependence between two variables, may thus be of at least two types. To represent a statistical dependence of an outcome on an explanatory feature, we use a directed edge with an arrow pointing to the outcome from the explanatory feature. For relations between features of equal standing, we use undirected edges.

In fact, it turns out to be useful to have two types of undirected edge. A dashed line is used to represent the dependence between two outcomes or responses given variables in their past. By contrast, a full line in the block of variables describing the background or context of the study and early features of the individuals under study, represents a conditional dependence given all remaining background variables.

From one viewpoint, the role of the graphical representation is to specify statistical independences that can be used to simplify understanding. From a complementary perspective, often the more immediately valuable, the purpose is to show those strong dependences that will be the base for interpreting pathways of dependence.

20.3 Some History of Graphical Markov Models

The development of graphical Markov models started with undirected, full line graphs; see Wermuth (1976), Darroch et al. (1980). The results built, for discrete random variables, on the log-linear models studied by Birch (1963), Goodman (1970), Bishop et al. (1975), and for Gaussian variables, on the covariance selection models by Dempster (1972). Shortly later, the models were extended to acyclic directed graph models for Gaussian and for discrete random variables; see Wermuth (1980), Wermuth and Lauritzen (1983). With the new model classes, results from the beginning of the 20th century by geneticist Sewall Wright and by probabilist Andrej Markov were combined and extended.

These generalizations differ from those achieved with structural equations that were studied intensively in the 1950s within econometrics; see, for instance, Bollen (1989). Structural equation models extend sequences of linear, multiple regression equations by permitting explicitly endogenous responses. These have residuals that are correlated with some or all of the regressors. For such endogenous responses, equation parameters need not measure conditional dependences, missing edges in graphs of structural equations need not correspond to any independence statement and no simple local modelling may be feasible. This contrasts with traceable regressions; see Sect. 20.4.1.

Wright had used directed acyclic graphs, that is graphs with only directed edges and no variables of equal standing, to represent linear generating processes. He developed ‘path analysis’ to judge whether such processes were well compatible with his data. Path analyses were recognized by Tukey (1954) to be fully ordered, also called ‘recursive’, sequences of linear multiple regressions in standardized variables.

With his approach, Wright was far ahead of his time, since, for example, formal statistical tests of goodness of fit were developed much later; see Wilks (1938). Conditions under which directed acyclic graphs represent independence structures for almost arbitrary types of random variables were studied later still; see Pearl (1988), Studený (2005).

One main objective of traceable regressions is to uncover graphical representations that lead to an understanding of data generating processes. These are not restricted to linear relations although they may include linear processes as special cases. A probabilistic data generating process is a recursive sequence of conditional distributions in which response variables can be vector variables that may contain discrete or continuous components or both types. Each of the conditional distributions specifies both the dependences of a joint response, Y_a say, on components in an explanatory variable vector, Y_b, and the undirected dependences among individual response component pairs of Y_a.

Graphical Markov models generalize sequences of single responses and single explanatory variables that have been extensively studied as Markov chains. Markov had recognized at the beginning of the 20th century that seemingly complex joint probability distributions may be radically simplified by using the notion of conditional independence.

In a Markov chain of random variables $Y_1, \ldots, Y_d$, the joint distribution is built up by starting with the marginal density f_d of Y_d and generating then the conditional density $f_{d-1|d}$. At the next step, conditional independence of Y_{d-2} from Y_d given Y_{d-1} is taken into account, with $f_{d-2|d-1,d} = f_{d-2|d-1}$. One continues such that with $f_{i|i+1,\ldots d} = f_{i|i+1}$, response Y_i is conditionally independent of $Y_{i+2}, \ldots, Y_d$ given Y_{i+1}, written compactly in terms of nodes as $i \perp\!\!\!\perp \{i+2, \ldots, d\} \mid \{i+1\}$, and ends, finally, with $f_{1|2,\ldots,d} = f_{1|2}$, where Y_1 has just Y_2 as an important, directly explanatory variable.

The fully directed graph, that captures such a Markov chain, is a single directed path of arrows. For five nodes, $d = 5$, and node set $N = \{1, 2, 3, 4, 5\}$, the graph is

$$1 \prec\!\!- 2 \prec\!\!- 3 \prec\!\!- 4 \prec\!\!- 5.$$

This graph corresponds to a factorization of the joint density f_N given by

$$f_N = f_{1|2} f_{2|3} f_{3|4} f_{4|5} f_5.$$

The three defining local independence statements given directly by the above factorization or by the graph are: $1 \perp\!\!\!\perp \{3, 4, 5\} \mid 2$, $2 \perp\!\!\!\perp \{4, 5\} \mid 3$ and $3 \perp\!\!\!\perp 5 \mid 4$. One also says that in such a generating process, each response Y_i 'remembers of its past just the nearest neighbour', the nearest past variable Y_{i+1}.

Directed acyclic graphs are the most direct generalization of Markov chains. They have a fully ordered sequence of single nodes, representing individual response variables for which conditional densities given their past generate f_N. No pairs of variables are on an equal standing. In contrast to a simple Markov chain, in this more general setting, each response may 'remember any subset or all of the variables in its past'.

Directed acyclic graphs are also used for Bayesian networks where the node set may not only consist of random variables, that correspond to features of observable units, but can represent decisions or parameters. As a framework for understanding possible causes and risk factors, directed acyclic are too limited since they exclude the possibility of an intervention affecting several responses simultaneously.

One early objective of graphical Markov models was to capture independence structures by appropriate graphs. As mentioned before, an independence structure is the set of all independence statements implied by the given graph. Such a structure is to be satisfied by any family of densities, f_N, said to be generated over a given graph.

In principle, all independence statements that arise from a given set of defining statements of a graph, may be derived from basic laws of probability by using the standard properties satisfied by any probability distribution and possibly some additional ones, as described for regression graphs in Sect. 20.4.1; see also Frydenberg (1990) for a discussion of properties needed to combine independence statements captured by directed acyclic graphs.

The above Markov chain implies for instance also

$$1 \perp\!\!\!\perp 4 \mid 3, \quad \{1,2\} \perp\!\!\!\perp \{4,5\} \mid 3, \quad \text{and} \quad 2 \perp\!\!\!\perp 4 \mid \{1,3,5\}.$$

For many variables, methods defined for graphs simplify considerably the task of deciding for a given independence statement whether it is implied by a graphs. Such methods have been called separation criteria; see Geiger et al. (1990), Lauritzen et al. (1990) and Marchetti and Wermuth (2009) for different but equivalent separation criteria for directed acyclic graphs.

For ordered sequences of vector variables, permitting joint instead of only single responses, the graphs are directed acyclic in blocks of vector variables. These blocks are sometimes called the ‘chain elements’ of the corresponding ‘chain graphs’. Four different types of such graphs for discrete variables have been classified and studied by Drton (2009). He proves that two types of chain graph have the desirable property of defining always curved exponential families for discrete distributions; see for instance Cox (2006) for the latter concept.

This property holds for the ‘LWF-chain graphs’ of Lauritzen and Wermuth (1989) and Frydenberg (1990), and for the graphs of Cox and Wermuth (1993, 1996) that have more recently been slightly extended and studied as ‘regression graphs’; see Wermuth and Sadeghi (2012), Sadeghi and Marchetti (2012). With the added feature that each edge in the graph corresponds to dependence that is substantial in a given context, they become ‘traceable regressions’; see Wermuth (2012).

Most books by statisticians on graphical Markov models focus on undirected graphs and on LWF-chain graphs; see Højsgaard et al. (2012), Edwards (2000), Lauritzen (1996), Whittaker (1990). In this latter class of graphical Markov models, each dependence between a response and a variable in its past is considered to be conditional also on all other components within the same joint response.

Main distinguishing features between different types of chain graph are the conditioning sets for the independences, associated with the missing edges, and for the edges present in the graph. For regression graphs, conditioning sets are always excluding other components of a given response vector, and criteria, to read off the graph all implied independences, do not change when the last chain element contains an undirected, full-line graph. It is in this general form, in which we introduce this class of models here. The separation criteria for these models are generalized versions of the criteria that apply to directed acyclic graphs.

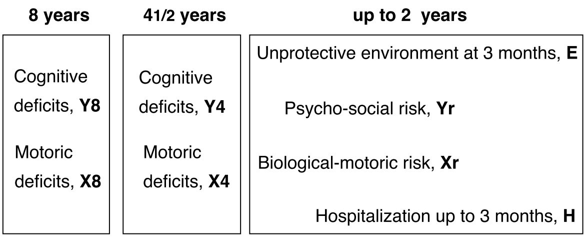

8 years
41/2 years
up to 2 years
Cognitive deficits, **Y8**
Motoric deficits, **X8**
Cognitive deficits, **Y4**
Motoric deficits, **X4**
Unprotective environment at 3 months, **E**
Psycho-social risk, **Yr**
Biological-motoric risk, **Xr**
Hospitalization up to 3 months, **H**

Fig. 20.1 Ordering of the variables given by time; the joint responses of primary interest are Y_8, X_8, those of secondary interest are Y_4, X_4, the four context variables are risks known up to age 2

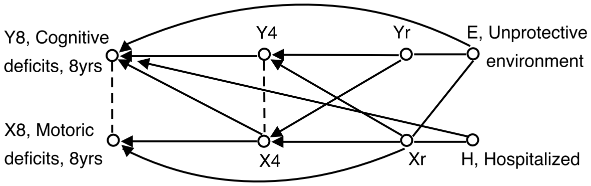


Fig. 20.2 A well-fitting regression graph for data of the child development study; *arrows* pointing from regressors in the past to a response in the future; *dashed lines* for dependent responses given their past; *full lines* for dependent early risk factors given the remaining background variables

Figure 20.1 shows two sets of joint responses and a set of background variables, ordered by time. The two related joint responses concern aspects of cognitive and motor development at age 8 years (abbreviated by Y_8, X_8, respectively) and at age 4.5 years (Y_4, X_4). There are two risks, measured up to 2 years, Y_r, X_r, where Y_r is regarded as a main risk for cognitive development and X_r as a main risk for motor development. Two more potential risks are available already at age 3 months of the child. Detailed definitions of the variables, a description of the study design and of further statistical results are given in Laucht et al. (1997) and summarized in Wermuth and Laucht (2012).

20.4 Sequences of Regressions and Their Regression Graphs

The well-fitting regression graph in Fig. 20.2 is for the variables of Fig. 20.1 and for data of 347 families participating in the Mannheim study from birth of their first child until the child reached the age of 8 years. The graph results from the statistical analyses reported in Sect. 20.4.2. These are further discussed in Sect. 20.4.3.

The goodness-of-fit of the graph to the given data is assessed by local modeling which include here linear and nonlinear dependences. The following Table 20.1 gives a summary in terms of Wilkinson's model notation that is in common use for generalized linear models and two coefficients of determination, R^2. There is a good fit for quantitative responses when the changes from R^2_{full} to R^2_{sel} are small, that is from the regression of an individual response on all variables in its past to a regression on only a reduced set of selected regressors.

Table 20.1 Fitted equations in Wilkinson's notation. Note that any square of a variable implies that also its main effect is included

Response	Selected model	R^2_{full}	R^2_{sel}
Y_8:	$Y_4 + X_4^2 + E + H$	0.67	0.67
X_8:	$X_4^2 + X_r$	0.36	0.36
Y_4:	$Y_r + X_r^2$	0.25	0.25
X_4:	$Y_r + X_r^2$	0.37	0.36
Y_r:	E^2	0.57	0.56
X_r:	$E + H$	0.35	0.35

20.4.1 Explanations and Definitions

In each regression graph, arrows point from the past to the future. An arrow is present, between a response and a variable in its past, when there is a substantively important dependence, that is also statistically significant, given all its remaining regressors. Regressors are recognized in the graph by arrows pointing to a given response node.

The undirected dependence between two individual components of a response vector is indicated here by a dashed line; some authors draw instead a bi-directed edge. Such an edge is present if there is a substantial dependence between two response components given the past of the considered joint response. An undirected edge between two context variables is a full line. Such an edge is present when there is a substantial dependence given the remaining context variables. An edge is missing, when for this variable pair no dependence can be detected, of the type just described.

The important elements of this representation are node pairs i, k, possibly connected by an edge, and a full set ordering $g_1 < g_2 < \cdots < g_J$ for the connected components g_j of a regression graph. The connected components of the graph are uniquely obtained by deleting all arrows from the graph and keeping all nodes and all undirected edges. In general, several orderings may be compatible with a given graph since different generating processes may lead to a same independence structure.

There is further an ordered partitioning of the node set into two parts, that is a split of N as $N = (u, v)$, such that response node sets $g_1, \ldots$ are in u and background node sets $\ldots, g_J$ are in v. In Fig. 20.2, there are two sets in u: $g_1 = \{Y_8, X_8\}$ and $g_2 = \{Y_4, X_4\}$. The subgraph of the background variables is for $v = g_3 = \{Y_r, X_r, E, H\}$ and there is only one compatible ordering of the three sets g_j.

Within v, the undirected graph is commonly called a concentration graph, reminding us of the parameterization for a Gaussian distribution, where a concentration, an element in the inverse covariance matrix, is a multiple of the partial correlation given all remaining variables; see Cox and Wermuth (1996, Sect. 3.4) or Wermuth (1976).

Within u, the undirected graph induced by the set g_j is instead a conditional covariance graph given the past of g_j, the nodes in $g_{>j} = \{g_{j+1}, \ldots, g_J\}$; see

Wermuth et al. (2009), Wiedenbeck and Wermuth (2010) for related estimation tasks. Arrows may point from any node in g_j for $j > 1$ to its future in $g_{<j} = \{g_1, \ldots, g_{j-1}\}$ but never to its past. Thus within each g_j, there are only undirected edges and all arrows point from nodes in g_j to nodes in $g_{<j}$, where $g_{<1} = \emptyset$.

With $g_{>J} = \emptyset$, the basic factorization of a family of densities f_N, generated over a regression graph, G^N_{reg}, is

$$f_N = f_{u|v} f_v \quad \text{with } f_{u|v} = \prod_{g_j \subseteq u} f_{g_j | g_{>j}} \text{ and } f_v = \prod_{g_j \subseteq v} f_{g_j}, \tag{20.1}$$

and the family satisfies all independence constraints implied by the graph.

For i, k a node pair, and $c \subset N \setminus \{i, k\}$, we write $i \perp\!\!\!\perp k \mid c$ for Y_i, Y_k conditionally independent given Y_c. In terms of a joint conditional density $f_{ik|c}$, this is equivalent to the following constraints on conditional densities:

$$i \perp\!\!\!\perp k \mid c \iff (f_{i|kc} = f_{i|c}) \iff f_{ik|c} = (f_{i|c} f_{k|c}).$$

For every variable pair Y_i, Y_k making an important contribution to the generating process of f_N, we say it is conditionally dependent given Y_c for some $c \subset N \setminus \{i, k\}$ specified in Definition 20.1 below and write $i \pitchfork k \mid c$. A regression graph is said to be edge-minimal if every missing edge in the graph corresponds to a conditional independence statement and every edge present is taken to represent a dependence; see the following definition.

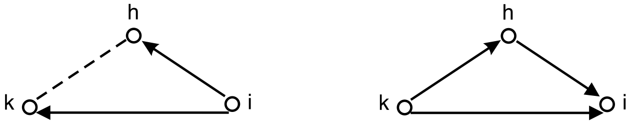
Definition 20.1 *Defining pairwise dependences of* G^N_{reg}. An edge-minimal regression graph specifies with $g_1 < \cdots < g_J$ a generating process for f_N where the following dependences

$$\begin{aligned} &i \;{-}\,{-}\,{-}\; k : \; i \pitchfork k \mid g_{>j} && \text{for } i, k \text{ response nodes in } g_j \text{ of } u, \\ &i \prec\!\!\!-\!\!\!-\; k : \; i \pitchfork k \mid g_{>j} \setminus \{k\} && \text{for response node } i \text{ in } g_j \text{ of } u \text{ and} \\ & && \text{node } k \text{ in } g_{>j}, \\ &i \;\text{——}\; k : \; i \pitchfork k \mid v \setminus \{i, k\} && \text{for } i, k \text{ context nodes in } v, \end{aligned} \tag{20.2}$$

define the edges present in G^N_{reg}. The meaning of each corresponding edge missing in G^N_{reg}results with the dependence sign $\pitchfork$ replaced by the independence sign$\perp\!\!\!\perp$.

By (20.2), a unique independence statement is assigned to the missing edge of each uncoupled node pair i, k. To combine independence statements implied by a regression graph, two properties are needed, called composition and intersection; see Sadeghi and Lauritzen (2013). The properties are stated below in Definition 20.3(1) as a same joint independence implied by the two independence statements under bullet points 2 and 3 on the right-hand side. In their simplest form, the two properties can be illustrated with two simple 3-node graphs.

For all trivariate probability distributions, one knows $i \perp\!\!\!\perp hk \implies (i \perp\!\!\!\perp h$ and $i \perp\!\!\!\perp k)$ as well as $i \perp\!\!\!\perp hk \implies (i \perp\!\!\!\perp h \mid k$ and $i \perp\!\!\!\perp k \mid h)$. The reverse implications are the composition and the intersection property, respectively. Thus, whenever node i is isolated from the coupled nodes h, k in a 3-node regression graph, it is to be interpreted as $i \perp\!\!\!\perp hk$ and this type of subgraph in three nodes i, h, k results, under composition, by removing the ih-arrow and the ik-arrow in the following graph on the left and under intersection in the following graph on the right. These small examples show already that the two properties are used implicitly in the selection of regressors; the composition property for multivariate regressions and the intersection property for directed acyclic graph models.



For the tracing of dependences, we need both of these properties but also the following, called singleton transitivity. It is best explained in terms of the Vs of a regression graph, the subgraphs in 3 nodes having 2 edges. In a regression graph, there can be at most 8 different V-configurations. Such a V in three nodes, (i, o, k) say, has uncoupled endpoints i, k and inner node o.

The V configurations in G^N_{reg} are of two different types. In G^N_{reg}, the collision Vs are:

$$i \; {-}\,{-}\,{-} \circ \prec\!\!\!\!\longleftarrow k, \quad i \longrightarrow\!\!\!\!\succ \circ \prec\!\!\!\!\longleftarrow k, \quad i \; {-}\,{-}\,{-} \circ {-}\,{-}\,{-} \; k,$$

and the transmitting Vs are:

$$i \prec\!\!\!\!\longleftarrow \circ \prec\!\!\!\!\longleftarrow k, \quad i \prec\!\!\!\!\longleftarrow \circ \longrightarrow k, \quad i \longrightarrow \circ \longrightarrow k, \quad i \prec\!\!\!\!\longleftarrow \circ {-}\,{-}\,{-} \; k, \quad i \prec\!\!\!\!\longleftarrow \circ \longrightarrow\!\!\!\!\succ k\,.$$

These generalize the 3 different possible Vs in a directed-acyclic graph. For such an edge-minimal graph, the two uncoupled nodes i, k of a transmitting V have either an important common-source node (as above on the right) or an important intermediate node (as above on the left), while the two uncoupled nodes i, k of a collision V with two arrows pointing to its inner node have an important, common response.

Singleton transitivity means that a unique independence statements is assigned to the endpoints i, k of each V of an edge-minimal graph, either the inner node o is included or excluded in every independence statement implied by the graph for i, k. For the strange parametrisation under which singleton transitivity is violated in a trivariate discrete family of distributions; see Wermuth (2012).

Expressed equivalently, let node pair i, k be uncoupled in an edge-minimal $G^N_{\text{reg}}$and consider a further node o and a set $c \subseteq N \setminus \{i, j, \text{o}\}$. Under singleton transitivity, for both the independences $i \perp\!\!\!\perp k \mid c$ and $i \perp\!\!\!\perp k \mid \text{o}c$ to hold, one of the constraints $\text{o} \perp\!\!\!\perp i \mid c$ or $\text{o} \perp\!\!\!\perp k \mid c$ has to be satisfied as well. Without singleton transitivity, the path of a V in nodes (i, o, k) can never induce a dependence for the endpoints i, k.

Definition 20.2 (Dependence-Base Regression Graph) An edge-minimal G^N_{reg}, is said to form a dependence base when its defining independences and dependences are combined by using standard properties of all probability distributions and the three additional properties: intersection, composition and singleton transitivity.

A dependence base regression graph, G^N_{reg}, is edge-inducing by marginalizing over the inner node of a transmitting V and by conditioning on the inner node of a collision V. This can be expressed more precisely.

Theorem 20.1 (Implications of Vs in a Dependence-Base Regression Graph (Wermuth 2012)) *For each* V *in three nodes*, (i, o, k) *of a dependence-base* G^N_{reg}, *there exists some* $c \subseteq N \setminus \{i, \text{o}, k\}$, *such that the graph implies* $(i \perp\!\!\!\perp k \mid \text{o}c$ *and* $i \pitchfork k \mid c)$ *when it is a transmitting* V, *while it implies* $(i \perp\!\!\!\perp k \mid c$ *and* $i \pitchfork k \mid \text{o}c)$ *when it is a collision* V.

The requirement appears to be elementary, but some densities or families of densities f_N, even when generated over a dependence base G^N_{reg}, may have such peculiar parameterizations that both statements $i \perp\!\!\!\perp k \mid \text{o}c$ and $i \perp\!\!\!\perp k \mid c$ can hold even though both node pairs i, o and o, k are coupled by an edge. Thus, singleton-transitivity needs to be explicitly carried over to a generated density.

We sum up as follows. For a successful tracing of pathways of dependence in an edge-minimal regression graph, all three properties are needed: composition, intersection and singleton transitivity. Intersection holds in all positive distributions and the composition property holds whenever nonlinear and interactive effects also have non-vanishing linear dependences or main effects.

Singleton transitivity is satisfied in binary distributions; see Simpson (1951). More generally, it holds when families of densities are generated over G^N_{reg}that have a rich enough parametrization, such as the conditional Gaussian distributions of Lauritzen and Wermuth (1989) that contain discrete and continuous responses.

Definition 20.3 *Characterizing properties of traceable regressions.* Traceable regression are densities f_N generated over a dependence base G^N_{reg}, that have for disjoint subsets a, b, c, d of N

(1) three equivalent decompositions of the same joint independence

- $b \perp\!\!\!\perp ac \mid d \iff (b \perp\!\!\!\perp a \mid cd$ and $b \perp\!\!\!\perp c \mid d)$,
- $b \perp\!\!\!\perp ac \mid d \iff (b \perp\!\!\!\perp a \mid d$ and $b \perp\!\!\!\perp c \mid d)$,
- $b \perp\!\!\!\perp ac \mid d \iff (b \perp\!\!\!\perp a \mid cd$ and $b \perp\!\!\!\perp c \mid ad)$, and

(2) edge-inducing V's of $G^N_{\text{reg}}$are dependence-inducing for f_N.

One outstanding feature of traceable regressions is that many of their consequences can be derived by just using the graph, for instance, when one is marginalizing over some variables in set M, and conditioning on other variables in set C. In particular, graphs can be obtained for node sets $N' = N \setminus \{C, M\}$ which capture precisely the independence structure implied by G^N_{reg}, the generating graph in the larger node set N, for $f_{N'|C}$, the family of densities of Y'_N given Y_C.

Such graphs are named independence-preserving, when they can be used to derive the independence structure that would have resulted from the generating graph by conditioning on a larger node set $\{C, c\}$ and marginalizing over the set $\{M, m\}$. Otherwise, such graphs are said to be only independence-predicting. Both types of graph transformations can be based on operators for binary matrices that represent graphs; see Wermuth et al. (2006), Wermuth and Cox (2004).

From a given generating graph, three corresponding types of independence-preserving graph result by using the same sets C, M. These are in a subclass of the much larger class of MC-graphs of Koster (2002), studied as the ribbon-less graphs by Sadeghi (2013a), or they are the maximal ancestral graphs of Richardson and Spirtes (2002) or the summary graphs of Wermuth (2011); see Sadeghi (2013a) for proofs of their Markov equivalence.

A summary graph shows when a generating conditional dependence, of Y_i on Y_k say, in f_N remains undistorted in $f_{N'|C}$, parametrized in terms of conditional dependences, and when it be may become severely distorted; see Wermuth and Cox (2008). Some of such distortions can occur in randomized intervention studies, but they may often be avoided by changing the set M or the set C.

Therefore, these induced graphs are relevant for the planning stage of follow-up studies, designed to replicate some of the results of a given large study by using a subset of the variables, that is after marginalizing over some variables, and/or by studying a subpopulation, that is after conditioning on another set of variables.

For marginalizing alone, that is in the case of $C = \emptyset$, one may apply the following rules for inserting edges repeatedly, keep only one of several induced edges of the same type, and gets often again a regression graph induced by $N' = N \setminus M$. In general, a summary graph results; see Wermuth (2011). The five transmitting Vs induce edges by marginalizing over the inner node

$$i \longleftarrow \oslash \longleftarrow k, \quad i \longleftarrow \oslash \text{———} k, \quad i \text{———} \oslash \text{———} k, \quad i \longleftarrow \oslash \,-\,-\,-\, k, \quad i \longleftarrow \oslash \longrightarrow k$$

to give, respectively,

$$i \longleftarrow k, \quad i \longleftarrow k, \quad i \text{———} k, \quad i \,-\,-\,-\, k, \quad i \,-\,-\,-\, k.$$

The induced edges 'remember the type of edge at the endpoints of the V' when one takes into account that each edge $\circ \,-\,-\,-\, \circ$ in G^N_{reg}can be generated by a larger graph, that contains $\circ \longleftarrow \oslash \longrightarrow \circ$. Thereby, the independence structure implied by this graph, for the node set excluding the hidden nodes, $\{\oslash\}$, is unchanged.

For any choice of C, M and a given generating graph G^N_{reg}, routines in the package 'ggm', contained within the computing environment R, help to derive the implications for $f_{N'|C}$ by computing either one of the different types of independence-preserving graph; see Sadeghi and Marchetti (2012). Other routines in 'ggm' decide whether a given independence-preserving graph is Markov equivalent to another one or to a graph in one of the subfamilies, such as a concentration or a directed acyclic graph; see Sadeghi (2013b) for justifications of these procedures. This helps to contemplate and judge possible alternative interpretations of a given G^N_{reg}.

For two regression graphs, the Markov equivalence criterion is especially simple: the two graphs have to have identical sets of node pairs with a collision V; see Theorem 1 of Wermuth and Sadeghi (2012). The result implies that the two sets may contain different ones of the 3 possible collision Vs. Also, the two sets of pairs with a transmitting V are then identical, though a given transmitting V in one graph may correspond in the other graph to another one of the 5 transmitting Vs that can occur in G^N_{reg}.

20.4.2 Constructing the Regression Graph via Statistical Analyses

As mentioned before, we use here data from the Mannheim Study of Children at risk. The study started in 1986 with a random sample of more than 100 newborns from the general population of children born in the Rhine–Neckar region in Germany. This sample was completed to give equal subsamples, in each of the nine level combinations of two types of adversity, taken to be at levels 'no, moderate or high'. In other words, there was heavy oversampling of children at risk.

The recruiting of families stopped with about 40 children of each risk level combination and 362 children in the study. All measurements were reported in standardized form using the mean and standard deviation of the starting random sample, called here the norm group. Of the 362 German-speaking families who entered the study when their first, single child was born without malformations or any other severe handicap, 347 families participated still when their child reached the age of 8 years.

Two types of risks were considered, one relevant for cognitive the other for motor development. One main difference to previous analyses is that we averaged three assessments of each type of risk: taken at birth, at 3 months and at two years. This is justified in both cases by the six observed pairwise correlations being all nearly equal. The averaged scores, called 'Psycho-social risk up to 2 years', Y_r, and 'Biological-motoric risk up to 2 years', X_r, have smaller variability than the individual components. This points to a more reliable risk assessment and leads to clearly recognizable dependences, to the edges present in Fig. 20.2.

The regression equations may be read off Tables 20.2 to 20.7. For instance for Y_8, there are four regressors and one nonlinear dependence on X_4 with

$$\mathrm{E}_{\text{lin}}\big(Y_8 \mid \text{past of } Y_8\big) = 0.03 + 0.78Y_8 + (0.07 + 0.10X_4)X_4 + 0.11E + 0.12H.$$

The test results of Table 20.2 imply that the previous measurement of cognitive deficits at age 4 years, Y_4 is the most important regressor and that the next important dependence is nonlinear and on motoric deficits at 4 years, X_4^2.

For each individual response component of the continuous joint responses, the results of linear-least squares fittings are summarized in six tables. In each case, the

Table 20.2 Regression results for Y_8

Response: Y_8, cognitive deficits at 8 years

Explanatory variables	Starting model			Selected			Excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
constant	0.00	–	–	0.03	–	–	–
Y_4, cognitive deficits, 4.5 yrs	0.78	0.05	15.36	0.78	0.05	15.70	–
X_4, motoric deficits, 4.5 yrs	0.05	0.04	–	0.07	0.04	–	–
Y_r, psycho-social risk, 2 yrs	0.00	0.07	0.01	–	–	–	−0.13
X_r, biol.-motoric risk, 2 yrs	0.07	0.07	1.07	–	–	–	1.08
E, Unprotect. environm., 3 mths	0.10	0.06	1.81	0.12	0.04	2.62	–
H, Hospitalisation up to 3 mths	0.09	0.05	1.91	0.12	0.04	3.00	–
X_4^2	0.09	0.01	6.53	0.10	0.01	7.15	–

$R^2_{\text{full}} = 0.67$, Selected model $Y_8 : Y_4 + X_4^2 + E + H$, $R^2_{\text{sel}} = 0.67$

Table 20.3 Regression results for X_8

Response: X_8, motoric deficits at 8 years

Explanatory variables	Starting model			Selected			Excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
constant	0.26	–	0.26	–	–	–	–
Y_4, cognitive deficits, 4.5 yrs	−0.01	0.06	−0.10	–	–	–	0.04
X_4, motoric deficits, 4.5 yrs	0.33	0.04	7.39	0.33	0.04	–	–
Y_r, psycho-social risk, 2 yrs	0.01	0.08	0.19	–	–	–	0.43
X_r, biol.-motoric risk, 2 yrs	0.17	0.08	2.27	0.19	0.06	2.97	–
E, Unprotect. environm., 3 mths	0.01	0.07	0.17	–	–	–	0.44
H, Hospitalisation up to 3 mths	0.01	0.08	0.26	–	–	–	0.26
X_4^2	0.18	0.23	3.41	0.05	0.02	2.89	–

$R^2_{\text{full}} = 0.36$, Selected model $X_8 : X_4^2 + X_r$, $R^2_{\text{sel}} = 0.36$

Table 20.4 Regression results for Y_4

Response: Y_4, cognitive deficits at 4.5 years

Explanatory variables	Starting model			Selected			Excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
constant	−0.29	–	–	−0.29	–	–	–
Y_r, psycho-social risk, 2 yrs	0.36	0.08	4.81	0.36	0.05	6.77	–
X_r, biol.-motoric risk, 2 yrs	0.17	0.09	–	0.18	0.07	–	–
E, Unprotect. environm., 3 mths	−0.01	0.07	−0.14	–	–	–	0.39
H, Hospitalisation up to 3 mths	0.14	0.04	3.36	–	–	–	−0.12
X_r^2	0.14	0.04	3.36	0.14	0.04	3.36	–

$R^2_{\text{full}} = 0.25$, Selected model $Y_4 : Y_r + X_r^2$, $R^2_{\text{sel}} = 0.25$

Table 20.5 Regression results for X_4

Response: X_4, motoric deficits at 4.5 years

Explanatory variables	Starting model			Selected			Excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
constant	−0.47	–	–	−0.47	–	–	–
Y_r, psycho-social risk, 2 yrs	0.33	0.10	3.44	0.28	0.07	4.21	–
X_r, biol.-motoric risk, 2 yrs	0.62	0.11	5.50	0.50	0.09	–	–
E, Unprotect. environm., 3 mths	−0.06	0.08	−0.66	–	–	–	−0.77
H, Hospitalisation up to 3 mths	−0.13	0.07	−1.83	–	–	–	−1.88
$(X_r)^2$	0.21	0.05	3.97	0.23	0.05	4.43	–

$R^2_{\text{full}} = 0.37$, Selected model $X_4 : Y_r + X_r^2$, $R^2_{\text{sel}} = 0.36$

Table 20.6 Regression results for Y_r

Response: Y_r, psycho-social risk up to 2 years

Explanatory variables	Starting model			Selected			Excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
constant	−0.20	–	–	−0.21	–	–	–
X_r, biol.-motoric risk, 2 yrs	−0.04	0.04	−0.81	–	–	–	−1.51
E, Unprotect. environm., 3 mths	0.57	0.03	–	0.55	0.03	–	–
H, Hospitalisation up to 3 mths	−0.03	0.04	−0.80	–	–	–	−1.50
E^2	0.16	0.03	6.12	0.16	0.03	6.20	–

$R^2_{\text{full}} = 0.57$, Selected model $Y_r : E^2$, $R^2_{\text{sel}} = 0.56$

Table 20.7 Regression results for X_r

Response: X_r, biologic-motoric risk up to 2 years

Explanatory variables	Starting model			Selected			Excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
constant	0.25	–	–	0.22	–	–	–
Y_r, psycho-social risk, 2 yrs	−0.05	0.07	−0.81	–	–	–	−1.22
E, Unprotect. environm., 3 mths	0.17	0.06	3.04	0.12	0.04	–	–
H, Hospitalisation up to 3 mths	0.48	0.04	12.30	0.48	0.04	12.40	–
E^2	−0.04	0.03	−1.09	–	–	–	−1.42

$R^2_{\text{full}} = 0.35$, Selected model $X_r : E + H$, $R^2_{\text{sel}} = 0.35$

response is regressed in the starting model on all the variables in its past. Quadratic or interaction terms are included whenever there is a priori knowledge or a systematic screening alerts to them; see Cox and Wermuth (1994).

The tables give the estimated constant term and for each variable in the regression, its estimated coefficient (coeff), the estimated standard deviation of the coefficient (s_{coeff}), as well as the ratio $z_{\text{obs}} = \text{coeff}/s_{\text{coeff}}$, often called a studentized value. Each ratio is compared to the 0.995 quantile of a standard Gaussian random variable Z, for which $\Pr(Z > |2.58|) = 0.01$. This relatively strict criterion for excluding variables assures that each edge in the constructed regression graph corresponds to a dependence that is considered to be substantively strong in the given context, in addition to being statistically significant for the given sample size.

At each backward selection step, the variable with the smallest observed value $|z_{\text{obs}}|$ is deleted from the regression equation, one at a time, until the threshold is reached so that no more variables can be excluded. The remaining variables are selected as the regressors of the response. An arrow is added for each of the regressors to the graph containing just the nodes, arranged in $g_1 < g_2 < \cdots < g_J$.

The last column in each table shows the studentized value z'_{obs}, that would be obtained when the variable were included next into the selected regression equation. Wilkinson's model notation is added in the table to write the selected model in compact form. For continuous responses, the coefficient of determination is recorded for the starting model, denoted by R^2_{full} and for the reduced model containing the selected regressors, denoted by R^2_{sel}.

A dashed line is added, for a variable pair of a given joint response, when in the regression of one on the other, there is a significant dependence given their combined set of the previously selected regressors.

A full line is added for a variable pair among the background variables, when in the regression of one on all the remaining background variables, there is a significant dependence of this pair. This exploits that an undirected edge present in a concentration graph, must also be significant in such a regression; see Wermuth (1992).

This strategy leads to a well-fitting model, unless one of the excluded variables has a too large contribution when it is added alone to a set of selected regressors. Such a variable would have to be included as an additional regressor. However, this did not happen for the given set of data.

The tests for the residual dependence of the two response components gives a weak dependence at age 8 with $z_{\text{obs}} = 2.4$ but a strong dependence at age 4.5 with $z_{\text{obs}} = 7.0$.

A global goodness-of-fit test, with proper estimates under the full model, may depend on additional distributional assumptions and require iterative fitting procedures. For exclusively linear relations of a joint Gaussian distribution, such a global test for the joint regressions would be equivalent to the fitting of a corresponding structural equation model, given the unconstrained background variables, and the global fitting of the concentration graph model to the context variables would correspond to estimation and testing for one of Dempster's covariance selection models.

20.4.3 Using a Well-Fitting Graph

There are direct and indirect pathways from risks at three months to cognitive deficits at 8 years. The exclusively positive conditional dependences along different

Fig. 20.3 A graph equivalent to the one of Fig. 20.2 with one hidden, common explanatory variable

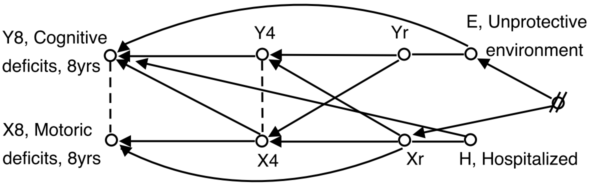
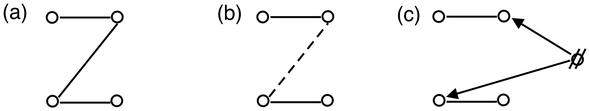


Fig. 20.4 A hidden variable graph (**c**) generating two Markov equivalent graphs (**a**) and (**b**)



paths accumulate to positive marginal dependences, even for responses connected only indirectly to a risk, for instance for Y_8 to Y_r or X_8 to E.

Among the background variables, an unprotective environment for the 3 months-old child, E, is strongly related to the psycho-social risk up to 2 years, Y_r and hospitalization up to 3 months, H, to the biological-motoric risk up to 2 years, X_r. The weakest but still statistically significant dependence among these four risks occurs for an unprotective environment, E, and the biological-motoric risk, X_r.

Such a dependence taken alone can often best be explained by an underlying common explanatory variable, here for instance a genetic or a socio-economic risk. This would lead to replacing the full line for (E, X_r) in Fig. 20.2 by the common-source V, shown in Fig. 20.3. The inner node of this V is crossed out because it represents a hidden that is unobserved variable. Hidden nodes represent variables that are unmeasured in a given study but whose relevance and existence is known or assumed.

Though Fig. 20.3 appears to contain only a small change compared to Fig. 20.2, this change requires a Markov equivalence result for a larger class than regression graphs, as available for the ribbon-less graphs of Sadeghi (2013a), since a path of the type i —— o ≺— k does not occur in a regression graph. Given these results, it follows that graphs Fig. 20.4(a) and (b) are Markov equivalent and that the structure of graph Fig. 20.4(b) can be generated by the larger graph Fig. 20.4(c) that includes a common, but hidden regressor node for the two inner nodes of the path.

To better understand the distinguishing features of the pathways of dependence in Fig. 20.2 leading to the joint responses of main interest at age 8, we generate the implied regressions graphs when the assessments at age 8 and at 4.5 years are available for only one of the two aspects. In that case one has ignored, that is marginalized over, the assessments of the other aspect at age 8 and 4.5.

The resulting graph, for Y_8 and Y_4 ignored, happens to coincide with the subgraph induced by the remaining, selected six nodes in Fig. 20.1, as shown in Fig. 20.5. Such an induced graph has the selected nodes and as edges all those present among them in the starting graph and no more. The graph of Fig. 20.5 implies that possible psycho-social risks of a child up to age 2, Y_r, do not contribute

Fig. 20.5 The regression graph induced by ignoring Y_8 and Y_4 in Figure 20.2; $M = \{Y_8, Y_4\}$, $C = \emptyset$

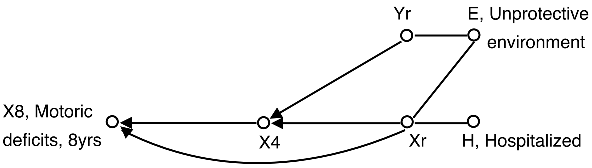
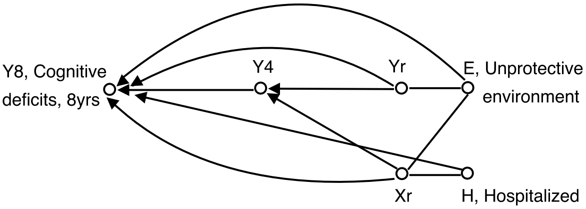


Fig. 20.6 The regression graph induced by ignoring X_8 and X_4 in Figure 20.2; $M = \{X_8, X_4\}$, $C = \emptyset$



directly to predicting motoric deficits at school-age, X_8, also when the more recent information on cognitive deficits is not available.

By contrast, the regression graph in Fig. 20.6 that results after ignoring X_8 and X_4, shows two additional arrows compared to the subgraph induced in Fig. 20.2 by Y_8, Y_4, Y_r, X_r, E, H.

The induced arrows are for (Y_8, Y_r) and for (Y_8, X_r). The graph suggests that cognitive deficits at school-age, Y_8, are directly dependent on all of the remaining variables when the more recent information on the motoric risks are unrecorded. There are direct and indirect pathways from H and from E to Y_8. They involve nonlinear dependences of cognitive deficits on previous motoric deficits or risks. These are recognized in the fitted equations but not directly in the graph alone.

What the graph also cannot show is that with X_8, X_4 unrecorded, the early risks, Y_r, H are less important as predictors when Y_4, X_r, X_r^2, E are available as regressors of Y_8. This effect is due to the strong partial dependences of Y_r, E^2 given E, X_r, H and of X_r, H given E, E^2, Y_r. Such implications, due to the special parametric constellations are not reflected in the graph alone.

Many more conclusions may be drawn by using just graphs like in Figs. 20.2 to 20.6. The substantive research questions and the special conditions of a given study are important; for some different types of study analyzed with graphical Markov models see, for instance, Klein et al. (1995), Gather et al. (2002), Hardt et al. (2004), Wermuth et al. (2012).

One major attraction of sequences of regressions in joint responses is that they may model longitudinal data from observational as well as from intervention studies. For instance, with fully randomized allocation of persons to a treatment, all arrows that may point to the treatment in an observational study, are removed from the regression graph. This removal reflects such a successful randomization: independence is assured for the treatment variable of all regressors or background variables, no matter whether they are observed or hidden.

20.5 Conclusions

The paper combines two main themes. One is the notion of traceable regressions. These are sequences of joint response regressions together with a set of background variables for which an associated regression graph not only captures an independence structure but permits the tracing of pathways of dependence. Study of such structures has both a long history and at the same time is the focus for much current development.

Joint responses are needed when causes or risk factors are expected to affect several responses simultaneously. Such situations occur frequently and cannot be adequately modeled with distributions generated over directed acyclic graph or such a graph with added dashed lines between responses and variables in their past to permit unmeasured confounders or endogenous responses.

A regression graph shows, in particular, conditional independences by missing edges and conditional dependences by edges present. The independences simplify the underlying data-generating process and emphasize the important dependences via the remaining edges. The dependences form the basis for interpretation, for the planning of or comparison with further studies and for possible policy action. Propagation of independencies is now reasonably well understood. There is scope for complementary further study that focuses on pathways of dependence.

The second theme concerns specific applications. Among the important issues here are an appropriate definition of population under study, especially when relatively rare events and conditions are to be investigated, appropriate sampling strategies, and the importance of building an understanding on step-by-step local analyses. The data of the Mannheim study happen to satisfy all properties needed for tracing pathways of dependence. This permits discussion of the advantages and limitations for some illustrated path tracings.

In the near future, more results on estimation and goodness of fit tests are to be expected, for instance by extending the fitting procedures for regression graph models of Marchetti and Lupparelli (2011) to mixtures of discrete and continuous variables, more results on the identification of models that include hidden variables such as those by Stanghellini and Vantaggi (2013) and those by Foygel et al. (2012), and further evaluations of properties of different types of parameters; see Xie et al. (2008) for an excellent starting discussion.

Acknowledgements The work by Nanny Wermuth reported in this paper was undertaken during her tenure of a Senior Visiting Scientist Award by the International Agency of Research on Cancer. We thank the referees, Bianca de Stavola and Rhian Daniel for their helpful comments. We used Matlab for statistical analyses.

References

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society. Series B. Methodological*, *25*, 220–233.

Bishop, Y. M. M., Fienberg, S. F., & Holland, P. W. (1975). *Discrete multivariate analysis*. Cambridge: MIT Press.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge: Cambridge University Press.

Cox, D. R., & Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statistical Science*, *8*, 204–218.

Cox, D. R., & Wermuth, N. (1994). Tests of linearity, multivariate normality and adequacy of linear scores. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *43*, 347–355.

Cox, D. R., & Wermuth, N. (1996). *Multivariate dependencies: models, analysis, and interpretation*. London: Chapman & Hall.

Darroch, J. N., Lauritzen, S. L., & Speed, T. P. (1980). Markov fields and log-linear models for contingency tables. *The Annals of Statistics*, *8*, 522–539.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, *28*, 157–175.

Drton, M. (2009). Discrete chain graph models. *Bernoulli*, *15*, 736–753.

Edwards, D. (2000). *Introduction to graphical modelling* (2nd ed.). New York: Springer.

Foygel, R., Draisma, J., & Drton, M. (2012). *Half-trek criterion for generic identifiability of linear structural equation models*. Submitted. doi:10.1214/12-AOS1012

Frydenberg, M. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics*, *17*, 333–353.

Gather, U., Imhoff, M., & Fried, R. (2002). Graphical models for multivariate time series from intensive care monitoring. *Statistics in Medicine*, *21*, 2685–2701.

Geiger, D., Verma, T. S., & Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks*, *20*, 507–534.

Goodman, L. A. (1970). The multivariate analysis of qualitative data: interaction among multiple classifications. *Journal of the American Statistical Association*, *65*, 226–256.

Hardt, J., Petrak, F., Filipas, D., & Egle, U. T. (2004). Adaptation to life after surgical removal of the bladder—an application of graphical Markov models for analysing longitudinal data. *Statistics in Medicine*, *23*, 649–666.

Højsgaard, S., Edwards, D., & Lauritzen, L. (2012). *Graphical Models with R*. Berlin: Springer.

Klein, J. P., Keiding, N., & Kreiner, S. (1995). Graphical models for panel studies, illustrated on data from the Framingham heart study. *Statistics in Medicine*, *14*, 1265–1290.

Koster, J. (2002). Marginalising and conditioning in graphical models. *Bernoulli*, *8*, 817–840.

Laucht, M., Esser, G., & Schmidt, M. H. (1997). Developmental outcome of infants born with biological and psychosocial risks. *Journal of Child Psychology and Psychiatry*, *38*, 843–853.

Lauritzen, S. L. (1996). *Graphical models*. Oxford: Oxford University Press.

Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, *17*, 31–57.

Lauritzen, S. L., Dawid, A. P., Larsen, B., & Leimer, H. G. (1990). Independence properties of directed Markov fields. *Networks*, *20*, 491–505.

Marchetti, G. M., & Lupparelli, M. (2011). Chain graph models of multivariate regression type for categorical data. *Bernoulli*, *17*, 827–844.

Marchetti, G. M., & Wermuth, N. (2009). Matrix representations and independencies in directed acyclic graphs. *The Annals of Statistics*, *47*, 961–978.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo: Morgan Kaufmann.

Richardson, T. S., & Spirtes, P. (2002). Ancestral Markov graphical models. *The Annals of Statistics*, *30*, 962–1030.

Sadeghi, K. (2013a, to appear). Representing modified independence structures. *Bernoulli*. arXiv:1110.4168

Sadeghi, K. (2013b). *Markov equivalences of subclasses of loopless mixed graphs*. Submitted. arXiv:1110.4539

Sadeghi, K., & Lauritzen, S. L. (2013, to appear). Markov properties of mixed graphs. *Bernoulli*. arXiv:1109.5909

Sadeghi, K., & Marchetti, G. M. (2012). Graphical Markov models with mixed graphs in R. *The R Journal*, *4*, 65–73.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B. Methodological*, *13*, 238–241.

Stanghellini, E., & Vantaggi, B. (2013, to appear). On the identification of discrete graphical models with hidden nodes. *Bernoulli*. doi:10.3150/12-BEJ435

Studený, M. (2005). *Probabilistic conditional independence structures*. London: Springer.

Tukey, J. W. (1954). Causation, regression, and path analysis. In O. Kempthorne, T. A. Bancroft, J. W. Gowen, & J. L. Lush (Eds.), *Statistics and mathematics in biology* (pp. 35–66). Ames: Iowa State University Press.

Wermuth, N. (1976). Analogies between multiplicative models for contingency tables and covariance selection. *Biometrics*, *32*, 95–108.

Wermuth, N. (1980). Linear recursive equations, covariance selection, and path analysis. *Journal of the American Statistical Association*, *75*, 963–997.

Wermuth, N. (1992). On block-recursive regression equations (with discussion). *Brazilian Journal of Probability and Statistics*, *6*, 1–56.

Wermuth, N. (2011). Probability models with summary graph structure. *Bernoulli*, *17*, 845–879.

Wermuth, N. (2012). Traceable regressions. *International Statistical Review*, *80*, 415–438.

Wermuth, N., & Cox, D. R. (2004). Joint response graphs and separation induced by triangular systems. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *66*, 687–717.

Wermuth, N., & Cox, D. R. (2008). Distortions of effects caused by indirect confounding. *Biometrika*, *95*, 17–33.

Wermuth, N., & Laucht, M. (2012). *Explaining developmental deficits of school-aged children*. Submitted.

Wermuth, N., & Lauritzen, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika*, *70*, 537–552.

Wermuth, N., & Sadeghi, K. (2012). Sequences of regressions and their independences (with discussion). *Test*, *21*, 215–279.

Wermuth, N., Wiedenbeck, M., & Cox, D. R. (2006). Partial inversion for linear systems and partial closure of independence graphs. *BIT*, *46*, 883–901.

Wermuth, N., Cox, D. R., & Marchetti, G. M. (2009). Triangular systems for symmetric binary variables. *Electronic Journal of Statistics*, *3*, 932–955.

Wermuth, N., Marchetti, G. M., & Byrnes, G. (2012). *Case-control studies for rare diseases: estimation of joint risks and of pathways of dependences*. Submitted.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Chichester: Wiley.

Wiedenbeck, M., & Wermuth, N. (2010). Changing parameters by partial mappings. *Statistica Sinica*, *20*, 823–836.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*, 60–62.

Xie, X. C., Ma, Z. M., & Geng, Z. (2008). Some association measures and their collapsibility. *Statistica Sinica*, *19*, 1165–1183.

Chapter 21
Data Mining in Pharmacoepidemiological Databases

Marc Suling, Robert Weber, and Iris Pigeot

21.1 Introduction

After market release of a newly developed drug, it needs to be systematically monitored in the post-marketing setting as not all possible safety risks can be uncovered during the clinical testing phase. The detection of possibly hazardous health outcomes of a drug usually relies on the application of signal detection methods (Edwards and Biriell 1994). Traditionally, these methods are applied to spontaneous reporting (SR) data, provided by health care professionals, patients or the pharmaceutical industry when an association between an exposure to a drug and the observed event is suspected (Hauben et al. 2005; Almenoff et al. 2005). However, secondary health-related data become more and more available for scientific use, such as electronic health records (EHRs) or claims data from health insurances. This allows to study a wide range of adverse drug effects. These secondary data usually cover substantially larger and broader populations than the SR data, reflect daily practice, and have longer follow-up periods, which makes it possible to successfully detect even rare events (Berlin et al. 2008; Vray et al. 2005; Vandenbroucke and Psaty 2008). Electronic health care databases are usually large in size, high-dimensional and of high complexity as they bear an unknown potential of interdependencies between the different variables. The size and the complexity of the databases make a manual analysis of every possible safety hazard impossible, and new sophisticated auto-

M. Suling (✉) · I. Pigeot
Leibniz Institute for Prevention Research and Epidemiology - BIPS, Achterstraße 30,
28359 Bremen, Germany
e-mail: suling@bips.uni-bremen.de

I. Pigeot
e-mail: pigeot@bips.uni-bremen.de

R. Weber
Oracle Deutschland B.V. & Co. KG, Schlossstraße 2, 13507 Berlin, Germany
e-mail: robert.weber@oracle.com

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_21, © Springer-Verlag Berlin Heidelberg 2013

Table 21.1 2 × 2 contingency table as basis for most signal detection techniques (DEC_{ij} = drug-event-combination of the exposure to drug i and the occurrence of adverse reaction j)

DEC_{ij}	Event	No event	Total
Exposed	n_{11}	n_{10}	$n_{1\cdot}$
Not exposed	n_{01}	n_{00}	$n_{0\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 0}$	$n_{\cdot\cdot}$

mated or semi-automated analysis techniques are needed to make beneficial use of these data in drug safety.

For SR data, several signal detection techniques have been developed that range from simple frequentistic ones like the proportional reporting ratio (PRR) or the reporting odds ratio (ROR) (Hauben et al. 2005) to more sophisticated Bayesian techniques like the Bayesian confidence propagation neural network (BCPNN) (Bate et al. 1998) or the multi-item Gamma-Poisson shrinker MGPS (DuMouchel 1999; DuMouchel and Pregibon 2001). These techniques can also be modified for the use on longitudinal data.

This chapter gives a brief introduction into the field of signal detection in drug safety monitoring and describes the most commonly used analysis techniques, both on SR data and on longitudinal health care data. We discuss a Bayesian technique which was applied to a German claims database to investigate its appropriateness for signal detection purposes. We describe the application of this Bayesian technique in a study on the adverse effects of phenprocoumon, compare the obtained results with the corresponding results of a case-control study, and briefly discuss our findings.

21.2 Methods

The most commonly used statistical analysis techniques for the detection of safety signals are the so-called disproportionality measures. They operate on frequencies of drug-event-combinations (DECs), collated in 2 × 2 contingency tables, as given in Table 21.1 (for applications in clinical settings, refer to the contribution by Wellmann, Chap. 22). For each DEC_{ij} (i.e., drug i, adverse drug reaction (ADR) j) in question, such a contingency table is constructed. Please note that for the sake of readability we omit the index ij for the number of event counts in Table 21.1 and in subsequent formulae wherever possible. In each of these tables, a disproportionality measure is calculated to determine if the respective DEC_{ij} in question is reported more often than one would expect, assuming independency of drug exposure and the occurrence of the event.

Although each single table is fairly simple, the sheer mass of possible combinations in large databases can make the analysis more complex. If one considers a database containing more than 15,000 drugs and more than 10,000 single events, a total of more than 150 million contingency tables need to be examined. In recent years, various disproportionality measures have been proposed, some of the most commonly used will be presented in detail below.

21.2.1 Frequentistic Risk Measures

The most basic and widely used frequentistic measures for disproportionality in such contingency tables are the ROR and the PRR (Rothman et al. 2004; Evans et al. 2001), defined as

$$\mathrm{ROR}_{ij} = \frac{P(\mathrm{ADR}\, j \mid \mathrm{drug}\, i)/P(\mathrm{no\ ADR}\, j \mid \mathrm{drug}\, i)}{P(\mathrm{ADR}\, j \mid \mathrm{no\ drug}\, i)/P(\mathrm{no\ ADR}\, j \mid \mathrm{no\ drug}\, i)} \tag{21.1}$$

and

$$\mathrm{PRR}_{ij} = \frac{P(\mathrm{ADR}\, j \mid \mathrm{drug}\, i)}{P(\mathrm{ADR}\, j \mid \mathrm{no\ drug}\, i)} \tag{21.2}$$

with $P(\mathrm{ADR}\, j \mid \mathrm{drug}\, i)$ denoting the probability of a report on the target adverse event j given the exposure to the target drug i. These measures are defined analogously to the odds ratio (OR) and relative risk (RR) that are the most common risk estimates used in epidemiological studies. The maximum-likelihood (ML) estimators of (21.1) and (21.2) are given as

$$\widehat{\mathrm{ROR}}_{ij} = \frac{n_{11}/n_{10}}{n_{01}/n_{00}}, \qquad \widehat{\mathrm{PRR}}_{ij} = \frac{n_{11}/n_{1\cdot}}{n_{01}/n_{0\cdot}}$$

(Van Puijenbroek et al. 2002), that are consistent and asymptotically normally distributed. Analogously to the approximate 95 % confidence intervals (CIs) for OR and RR (Morris and Gardner 1988), the approximate CIs for ROR and PRR can be obtained via the asymptotic normal distribution of their log-transformed estimators (Van Puijenbroek et al. 2002) as

$$\mathrm{CI}^{\mathrm{ROR}}_{\mathrm{lower/upper}} = \widehat{\mathrm{ROR}} \cdot e^{\pm 1.96 \cdot \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}}}$$

and

$$\mathrm{CI}^{\mathrm{PRR}}_{\mathrm{lower/upper}} = \widehat{\mathrm{PRR}} \cdot e^{\pm 1.96 \cdot \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{01}} - \left(\frac{1}{n_{1\cdot}} + \frac{1}{n_{0\cdot}}\right)}}.$$

The estimates of ROR and PRR tend to become statistically instable when a small number of events is observed. This may lead to large estimates with wide confidence intervals (Shibata and Hauben 2011; DuMouchel 1999), creating many false-positive signals for very rare events. Despite this shortcoming, these two methods are widely used for the detection of safety signals on SR data as they are easy to implement and do not need much computing time.

To overcome the instability of the above estimators when applied to small numbers of drug-event combinations, more advanced Bayesian shrinkage techniques have been developed in the recent years. The two methods mainly used today are the BCPNN, applied at the Uppsala Monitoring Centre (UMC) to analyze the SR database of the World Health Organization (WHO), and the MGPS, which is based on the Gamma-Poisson shrinker (GPS) by DuMouchel (1999) and deployed on the

SR data of the Food and Drug Administration (FDA) in the US. Both methods are based on the relative reporting ratio (RRR) defined as

$$\mathrm{RRR} = \frac{P(\mathrm{drug}\, i, \mathrm{ADR}\, j)}{P(\mathrm{drug}\, i) \cdot P(\mathrm{ADR}\, j)}$$

with $P(\mathrm{drug}\, i, \mathrm{ADR}\, j)$ denoting the joint probability of exposure to drug i and occurrence of adverse drug reaction j and $P(\mathrm{drug}\, i)$ as well as $P(\mathrm{ADR}\, j)$ the respective marginal probabilities. In a frequentistic approach, the ML-estimator of RRR reads as

$$\widehat{\mathrm{RRR}} = \frac{n_{11} \cdot n_{..}}{n_{1.} \cdot n_{.1}}. \tag{21.3}$$

21.2.2 Bayesian Shrinkage—The Gamma-Poisson Shrinker

The approach proposed by DuMouchel, the so-called Gamma-Poisson shrinker (GPS) algorithm (DuMouchel 1999) considers the occurrence of the target DEC as rare event, for which typically a Poisson distribution is assumed to hold. It is then of interest to investigate the relative reporting rate λ_{11} with

$$\lambda_{11} = \frac{\mu_{11}}{E_{11}},$$

where μ_{11} is the mean of the Poisson distribution of the observed DEC count n_{11} and E_{11} denotes the expected event count under the assumption that drug exposure i and ADR j are independent.

Following an empirical Bayes approach now, λ_{11} itself is treated as random and not as fixed parameter, i.e., as realization of a random variable Λ_{11}. Thus, the number N_{11} of combinations of drug i and ADR j is assumed to be conditionally Poisson distributed given Λ_{11}, where Λ_{11} is assumed to be gamma distributed with density function $g(\lambda_{11} \mid \alpha, \beta), \alpha, \beta > 0$, mean $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}$. To enhance the flexibility, not a single gamma distribution but a mixture of two gamma distributions with initially unknown mixture parameter $p > 0$ is assumed as a-priori distribution Γ_{prior} of Λ_{11} with density function g_{prior}:

$$\begin{aligned} g_{\mathrm{prior}}\big(\lambda_{11} \mid p, \alpha_1, \alpha_2, \beta_1, \beta_2\big) &= p \cdot g\big(\lambda_{11} \mid \alpha_1, \beta_1\big) + (1-p) \cdot g\big(\lambda_{11} \mid \alpha_2, \beta_2\big) \\ &= p \cdot \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \lambda_{11}^{\alpha_1 - 1} \cdot e^{-\lambda_{11}\beta_1} \\ &\quad + (1-p) \cdot \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} \cdot \lambda_{11}^{\alpha_2 - 1} \cdot e^{-\lambda_{11}\beta_2}, \end{aligned}$$

where $\Gamma(\cdot)$ denotes the gamma function. Using the formula for the k-th moment of a gamma distributed random variable X,

$$\mathbb{E}\big(X^k\big) = \beta^{-k} \prod_{i=0}^{k-1} (\alpha + i),$$

it can easily be shown that the mean and the variance of Λ_{11} can be calculated as

$$\begin{aligned} \mathbb{E}(\Lambda_{11}) &= p \cdot \frac{\alpha_1}{\beta_1} + (1-p) \cdot \frac{\alpha_2}{\beta_2}, \\ \operatorname{Var}(\Lambda_{11}) &= p(1-p) \cdot \left(\frac{\alpha_1}{\beta_1} - \frac{\alpha_2}{\beta_2} \right)^2 \\ &\quad + p \cdot \frac{\alpha_1}{\beta_1^2} + (1-p) \cdot \frac{\alpha_2}{\beta_2^2}. \end{aligned}$$

To determine the posteriori distribution Γ_{post} of $(\Lambda_{11} \mid N_{11} = n_{11})$ one needs to obtain the marginal distribution of N_{11}. Based on the well-known property that the marginal distribution of a random variable X is a negative binomial distribution if $X \mid Y$ is a Poisson distributed random variable and Y is gamma distributed, it can be directly shown that the marginal distribution of N_{11} under the above assumptions of a mixture of two gamma distributions for Λ_{11} is also a mixture of two negative binomial distributions with

$$\begin{aligned} P(N_{11} = n_{11}) &= f\big(n_{11}, E_{11} \mid p, \alpha_1, \alpha_2, \beta_1, \beta_2\big) \\ &= p \cdot nb\big(n_{11}, E_{11} \mid \alpha_1, \beta_1\big) + (1-p) \cdot nb\big(n_{11}, E_{11} \mid \alpha_2, \beta_2\big) \end{aligned}$$

with probability density functions $nb(\cdot)$

$$nb\big(n_{11}, E_{11} \mid \alpha_l, \beta_l\big) = \frac{\Gamma(\alpha_l + n_{11})}{n_{11}!\Gamma(\alpha_l)} \cdot \left(1 + \frac{\beta_l}{E_{11}}\right)^{-n_{11}} \cdot \left(1 + \frac{E_{11}}{\beta_l}\right)^{-\alpha_l}, \quad l = 1, 2.$$

Since the gamma distribution is a conjugate prior, the posterior distribution of Λ_{11} can be calculated in a closed analytic form as a gamma distribution with modified parameters

$$\big(\Lambda_{11} \mid N_{11} = n_{11}\big) \sim \Gamma_{\text{post}}\big(q, \alpha_1^*, \alpha_2^*, \beta_1^*, \beta_2^*\big)$$

with

$$\alpha_l^* = \alpha_l + n_{11}, \qquad \beta_l^* = \beta_l + E_{11}, \quad l = 1, 2,$$

and

$$q = \frac{p \cdot nb(n_{11}, E_{11} \mid \alpha_1, \beta_1)}{p \cdot nb(n_{11}, E_{11} \mid \alpha_1, \beta_1) + (1-p) \cdot nb(n_{11}, E_{11} \mid \alpha_2, \beta_2)}, \tag{21.4}$$

where $0 \leq q \leq 1$ is the new mixture parameter. Thus, the posteriori density is given as

$$\begin{aligned} &g_{\text{post}}\bigl(\lambda_{11} \mid q, \alpha_1 + n_{11}, \alpha_2 + n_{11}, \beta_1 + E_{11}, \beta_2 + E_{11}\bigr) \\ &\quad = q \cdot g\bigl(\lambda_{11} \mid \alpha_1 + n_{11}, \beta_1 + E_{11}\bigr) + (1-q) \cdot g\bigl(\lambda_{11} \mid \alpha_2 + n_{11}, \beta_2 + E_{11}\bigr) \\ &\quad = q \cdot \left(\frac{\beta_1^{*\alpha_1^*}}{\Gamma(\alpha_1^*)} \cdot e^{-\lambda_{11}\beta_1^*} \cdot \lambda_{11}^{\alpha_1^*-1} \right) + (1-q) \cdot \left(\frac{\beta_2^{*\alpha_2^*}}{\Gamma(\alpha_2^*)} \cdot e^{-\lambda_{11}\beta_2^*} \cdot \lambda_{11}^{\alpha_2^*-1} \right). \end{aligned}$$

Based on the posterior distribution of Λ_{11}, the mean of $\log(\Lambda_{11})$ can be derived as

$$\begin{aligned} \mathbb{E}\bigl(\log(\Lambda_{11})\bigr) = q \cdot &\left[\Psi(\alpha_1 + n_{11}) - \log\left(\frac{1}{\beta_1} + E_{11} \right) \right] \\ &+ (1-q)\left[\Psi(\alpha_2 + n_{11}) - \log\left(\frac{1}{\beta_2} + E_{11} \right) \right], \end{aligned}$$

where $\Psi(\cdot)$ denotes the Digamma function and q is defined as in (21.4). The resulting risk measure, the so-called empirical Bayesian geometric mean (EBGM), is defined as

$$\text{EBGM} = e^{\mathbb{E}(\log(\Lambda_{11}))}. \tag{21.5}$$

The EBGM is estimated via a plug-in approach replacing the unknown parameters α_1, α_2, β_1, β_2 and p by their empirical Bayes estimators and the expected event count E_{11} by its estimator under the assumption of independence between the occurrence of drug i and ADR j:

$$\hat{E}_{11} = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{n_{\cdot\cdot}}. \tag{21.6}$$

The 5th percentile of the posterior distribution of Λ_{11}, denoted as "EB05", is interpreted as the lower one-sided 95 % confidence limit for EBGM, and, analogously, the 95th percentile as upper one-sided 95 % confidence limit.

This Bayesian estimator gives risk estimates quite similar to $\widehat{\text{RRR}}$ (21.3) when n_{11} is large, but leads to risk estimates that are more conservative when event counts are small, i.e., the risk estimates are considerably smaller and the confidence intervals less wide, hence the denomination "shrinkage estimate". While this shrinkage might obfuscate a real signal by reducing it to a non-conspicuous level, it helps to eliminate false-positive signals, which otherwise would have to be adjudicated subsequently.

The techniques discussed so far assess the risk of two-way DECs, i.e., one drug and one ADR. Another serious concern is due to potential interactions between several drugs taken simultaneously in relation to the occurrence of an ADR. A famous example is the interaction of cerivastatin and gemfibrozil, leading to an elevated risk of rhabdomyolysis and resulting in the withdrawal of cerivastatin from the worldwide market in 2001 (Furberg and Pitt 2001).

DuMouchel and Pregibon (2001) extended the GPS to deal with multi-item sets of a size $m > 2$ (e.g., $m = 3$; drug–drug-event interactions between drugs i, j and event k), and therefore called this method *multi-item* Gamma-Poisson shrinker (MGPS). From the respective 3-dimensional contingency table, in a first step, EBGM_{ijk} and E_{ijk} are estimated analogously to (21.5) and (21.6). In a second step, a log-linear model is fitted, the so-called "all-two-factor" model to estimate the expected frequency of the joint occurrence of both drugs and the event under the assumption that higher-level interactions than two-way interactions can be ignored. Based on this estimator, denoted as e_{All2F}, the authors define the EXCESS2 value as

$$\text{EXCESS2} = \left(\widehat{\text{EBGM}}_{ijk} \cdot \hat{E}_{ijk}\right) - e_{\text{All2F}}$$

with high EXCESS2 values of an examined triplet indicating a safety risk under joint exposure to both drugs.

21.2.3 Extension to Longitudinal Data

For the analysis of longitudinal observational data (e.g., claims data), one option is to convert the data structure to match the structure of SR data, so that the above mentioned techniques can be directly applied. An additional option is to modify the above estimators and algorithms to better fit the structure of longitudinal data and take full advantage of the available information.

A key information in longitudinal studies is the number of days a patient was under risk, i.e., the number of days the patient was exposed to the target drug. Let t_1 denote the total number of days the individuals in the data were under risk, t_0 the total number of days the individuals were observed without being under risk, n_{11} the number of ADRs j under exposure to drug i and n_{01} the number of ADRs j when not exposed to drug i.

Based on the frequentistic approach, we consider here, analogously to the RR or the PRR (21.2) for signal detection, the incidence rate ratio (more precisely: the incidence density ratio) which is heuristically defined as

$$\text{IRR}_{11} = \frac{\dfrac{\text{number of ADRs}\, j \text{ under exposure to drug}\, i}{\text{total person-time being under risk}}}{\dfrac{\text{number of ADRs}\, j \text{ not under exposure to drug}\, i}{\text{total person-time not being under risk}}}$$

to appropriately account for the person-time under observation. The incidence rate ratio can then directly be estimated as

$$\widehat{\text{IRR}}_{11} = \frac{n_{11}}{n_{01}} \cdot \frac{t_0}{t_1}. \tag{21.7}$$

The empirical Bayes approach can be extended to longitudinal data by a straightforward modification of the GSP algorithm (Schuemie 2011). We simply have to replace the estimator $\hat{E}_{11}$ according to (21.6) by the following estimator

$$\hat{E}_{11,\text{long}} = n_{01} \cdot \frac{t_1}{t_0}, \tag{21.8}$$

again accounting for the person-times when calculating the plug-in estimator of (21.5). This gives the "longitudinal" GPS (LGPS) algorithm as suggested by Schuemie. Please note that the LGPS is based on the assumption that the risk is time-invariant.

21.3 Application of a Bayesian Shrinkage Algorithm—Study on Bleeding Risk Under Phenprocoumon

Based on the idea of using automated quantitative signal detection and exploration for longitudinal health care data, the Health Sciences unit of Oracle Corp.®, Redwood Shores, CA, currently develop a modified version of the established MGPS-based "Empirica® Signal" (Oracle 2011) software. In the following, a prototype of this signal detection software, implementing classical as well as the new Bayesian algorithm suitable for mining electronic health records (EHR), is applied to the German Pharmacoepidemiological Research Database (GePaRD) (Pigeot and Ahrens 2008). In this software, two different estimation procedures are implemented: (a) a Bayesian risk measure that takes into account person-time, following the idea of the MGPS and similar to the approach by Schuemie (2011), briefly referred to as Bayesian relative risk (BRR), and (b) $\widehat{\text{IRR}}$ according to (21.7).

To assess the validity of the findings obtained from such an automated data mining procedure, all results are compared to the gold standard of fully adjusted risk estimates for the same drug-event combinations, derived from a case-control study by Behr et al. (2010a, 2010b). The risk estimate obtained from the signal detection tool is considered to be of "reasonable" size if it is covered by the corresponding 95 % confidence interval (CI) of the ORs derived from the respective case-control study.

We are interested in the risk of intracerebral hemorrhages (ICH) under exposure to phenprocoumon. As it is a well-known medical fact that anticoagulant compounds generally bear an increased hemorrhagic risk (Johnsen et al. 2003; Grønbæk et al. 2008; Sturgeon et al. 2007; Bamford et al. 1988; Mattle et al. 1989; Olsen et al. 2010), these risks should also be detected via a data mining tool in the pharmacoepidemiological database.

21.3.1 Results Obtained from the Data Mining Tool

The dataset used for the application of the signal detection tool contains data from 3,460,501 individuals. 63,215 (1.8 %) individuals are exposed to phenprocoumon

Table 21.2 Incidence rate ratios and Bayesian relative risk estimates obtained from the signal detection software in comparison to odds ratio estimates and 95 % confidence intervals derived from case-control studies on the risk of intracerebral hemorrhage under phenprocoumon exposure

Age group	$\widehat{\text{IRR}}$	95 % CI	BRR	95 % C	$\widehat{\text{OR}}$	95 % CI
All	11.1	[9.5;13.0]	3.3	[2.9;3.9]	–	–
<55	20.4	[11.2;37.1]	5.1	[2.7;8.7]	4.2	[2.4;7.2]
55–64	7.1	[4.8;10.4]	4.5	[3.1;6.5]	3.2	[2.4;4.2]
65–74	4.2	[3.2;5.4]	3.5	[2.7;4.5]	3.5	[3.0;4.1]
75–84	2.4	[1.9;3.2]	2.3	[1.7;2.9]	2.5	[2.2;2.9]
85+	3.4	[1.9;6.1]	2.5	[1.4;4.2]	2.4	[1.8;3.3]

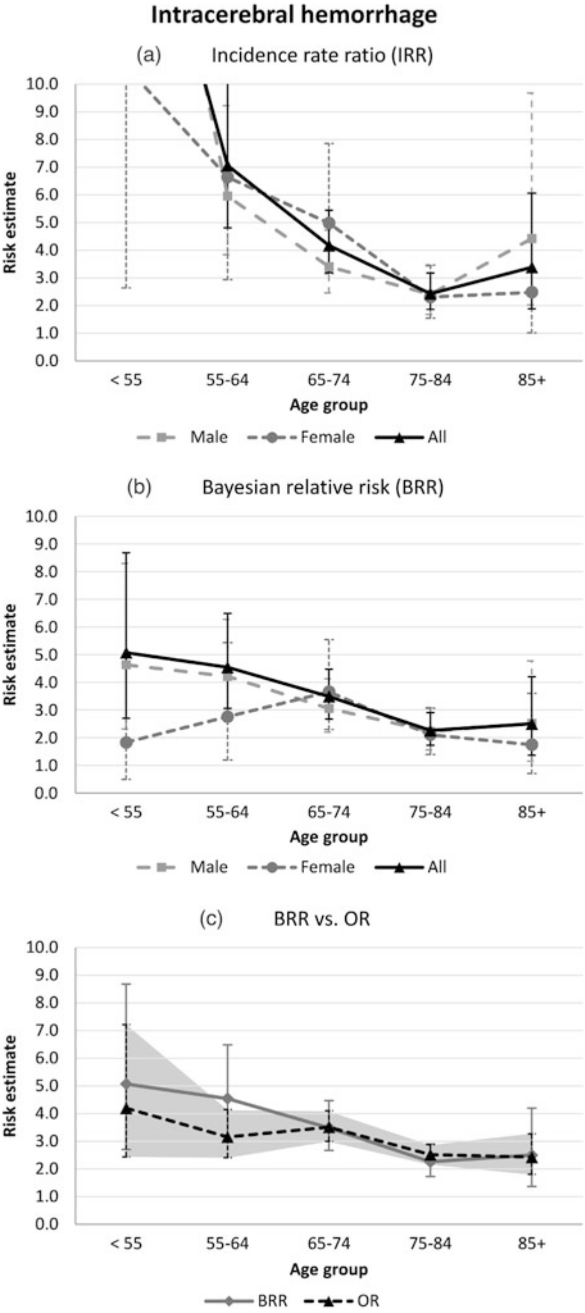
IRR = incidence rate ratio, BRR = Bayesian relative risk, OR = odds ratio, CI = confidence interval. $\widehat{\text{IRR}}$ and BRR are not adjusted for confounding besides stratification for age group and sex (results for stratification by sex not shown here)

during the study period and 2,301 individuals are diagnosed with ICH. 170 (7.4 %) of all 2,301 individuals with a diagnosis of ICH are exposed to phenprocoumon. For the overall hemorrhagic risk of phenprocoumon, the software yields an BRR estimate of 3.3 (95 % CI = [2.9; 3.9]). As expected, high IRR estimates are observed in the stratum <55 years, where both events and exposure are rare (event frequency and person-time not shown), i.e., $\widehat{\text{IRR}} = 20.4$ (95 % CI = [11.2; 37.1]). The BRR estimates tend to be smaller, especially in the age group <55. In general, IRR and BRR estimates become more similar with increasing age as illustrated in Figs. 21.1(a) and (b). In total, we find elevated risks under phenprocoumon exposure, as anticipated considering the findings reported in Behr et al. (2010a,b).

21.3.2 Comparison

When we compare the risk estimates obtained from the automated data mining tool with OR estimates obtained from the case-control studies by Behr et al. (cf. Table 21.2), results of the same size emerge for ICH. The BRR is always covered by the corresponding 95 % confidence interval of the fully adjusted OR estimate, except for patients with ICH aged 55–64, where the BRR of 4.5 (95 % CI = [3.1; 6.5]) exceeds the upper bound of the confidence interval of the OR (cf. Fig. 21.1(c)). Our results, based on a stratified data mining approach to determine the hemorrhagic risks of anticoagulant compounds, are of the same size as those obtained from the case-control studies. The estimated BRR is covered by the respective 95 % confidence intervals of the OR based on the case-control studies for most strata. As expected, $\widehat{\text{IRR}}$ tends to become statistically unstable in the lowest age group because of the small number of events and short duration under drug exposure compared to the huge number of unexposed days in that particular age group.

Fig. 21.1 Comparison of estimated risk measures for intracerebral hemorrhage under phenprocoumon exposure obtained from the signal detection software (incidence rate ratio (IRR) and Bayesian-adjusted relative risk (BRR)) and the case-control study (odds ratio (OR)). (**a**) $\widehat{\text{IRR}}$ and 95 % confidence intervals (CIs), stratified by sex and age group; (**b**) BRR and 95 % CIs, stratified by sex and age group; (**c**) BRR vs. $\widehat{\text{OR}}$ and 95 % CIs with CIs of the OR highlighted



21.4 Conclusions

The results from our application clearly endorse the feasibility of the automated data mining approach on pharmacoepidemiological databases like the GePaRD. We showed that in our example a signal known from literature can be successfully detected and that the risks found are also of the same size as risks obtained from case-control studies. However, one of the most crucial issues in observational studies as compared to clinical studies, the appropriate consideration of confounders in the statistical analysis, has not been addressed so far. In our study, $\widehat{\mathrm{IRR}}$ and, even more, BRR—both unadjusted except for age and sex—fit quite well to the fully adjusted $\widehat{\mathrm{OR}}$ from the case-control study. This suggests that a considerable amount of confounding was annihilated by stratification for sex and age. Nevertheless, further research on the effect of strong underlying confounding that cannot be controlled by simple stratification is surely needed.

One of the biggest problems regarding the adjustment for confounding in automated analyses is the selection of appropriate confounders. In pharmacoepidemiological studies this is typically done manually, which is not feasible in high-throughput automated analyses. Schneeweiss et al. (2009) proposed a technique for automated confounder selection based on propensity score (PS) matched cohorts (Rosenbaum and Rubin 1983; Austin 2011). The PS s_i of an individual i is defined as the probability to receive a certain treatment T_i, i.e. $T_i = 1$, given the observed vector of covariates $\mathbf{X}_i$, ignoring whether the respective subject actually received the treatment or not:

$$s_i = P\bigl(T_i = 1 \mid \mathbf{X}_i\bigr).$$

This simple definition allows for balancing the cohort underlying the analysis by matching treated to untreated subjects with the same PS. Thus, one obtains a cohort where the distributions of the observed covariates are balanced between both arms of the cohort. The high-dimensional propensity score (HDPS) algorithm proposed by Schneeweiss et al. automates the selection of the confounders to be considered in a multi-step approach. Briefly, the HDPS algorithm

(a) requires the identification of the different data dimensions (e.g., hospitalization data, outpatient care data, outpatient drug dispensation data) in the database,
(b) identifies a pre-specified number of the top most prevalent codes, e.g., ICD or ATC codes (ICD = international statistical classification of diseases and related health problems, ATC = anatomical therapeutic chemical classification system) in each data dimension as candidate covariates,
(c) ranks candidate covariates based on their recurrence (the frequency that the codes are recorded for each individual during the baseline period),
(d) ranks covariates across all data dimensions by their potential for control of confounding based on the bivariate associations of each covariate with the treatment and with the outcome,

(e) selects a pre-specified number of covariates from step 4 (e.g., 500) for PS modeling, and
(f) estimates the PS based on multivariable logistic regression using the selected covariates plus any pre-specified covariates.

This technique theoretically allows for a fully automated selection of confounders for each signal detection analysis. Rassen and Schneeweiss (2012) applied the HDPS to control for confounding in sequential database cohort studies. They concluded that HDPS offers substantial advantages over non-automated alternatives in active product safety monitoring systems.

We also applied the HDPS to the GePaRD to control for confounding in a study on the risk of upper gastrointestinal complications under exposure to non-steroidal anti-inflammatory drugs (Garbe et al. 2012). Since this turned out to be feasible with the structure of the database and successful with respect to an appropriate control for potential confounding, we consider the application of this algorithm to the German claims data in the signal detection context to be a worthwhile approach, although the computational run-time might still be too high for real-life drug monitoring.

Acknowledgements This study was conducted in the course of a project on hemorrhagic risk of anticoagulant compounds. The authors are grateful to all statutory health insurances that provided data for this project: the Allgemeine Ortskrankenkasse (AOK) Bremen/Bremerhaven, the Deutsche Angestellten-Krankenkasse (DAK—Unternehmen Leben), the Handelskrankenkasse (HKK) and the Techniker Krankenkasse (TK). Also the technical help provided by Robert Zambarano with the signal detection software from Oracle® Health Sciences is very much appreciated.

References

Almenoff, J. S., Tonning, J. M., Gould, A. L., Szarfman, A., Hauben, M., Ouellet-Hellstrom, R., Ball, R., Hornbuckle, K., Walsh, L., Yee, C., Sacks, S. T., Yuen, N., Patadia, V., Blum, M., Johnston, M., Gerrits, C., Seifert, H., & Lacroix, K. (2005). Perspectives on the use of data mining in pharmacovigilance. *Pharmacoepidemiology and Drug Safety*, *11*, 981–1007.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*, 399–424.

Bamford, J., Sandercock, P., Dennis, M., Warlow, C., Jones, L., McPherson, K., Vessey, M., Fowler, G., Molyneux, A., Hughes, T., Burn, J., & Wade, D. (1988). A prospective study of acute cerebrovascular disease in the community: the Oxfordshire community stroke project—1981-86. 2. Incidence, case fatality rates and overall outcome at one year of cerebral infarction, primary intracerebral and subarachnoid haemorrhage. *Journal of Neurology, Neurosurgery and Psychiatry*, *1*, 16–22.

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., & De Freitas, R. M. (1998). A Bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, *4*, 315–321.

Behr, S., Andersohn, F., & Garbe, E. (2010a). Risk of intracerebral hemorrhage associated with phenprocoumon exposure: a nested case-control study in a large population-based German database. *Pharmacoepidemiology and Drug Safety*, *7*, 722–730.

Behr, S., Andersohn, F., & Garbe, E. (2010b). Phenprocoumon exposure and risk of intracerebral bleeding. In *Abstracts of the 26th international conference on pharmacoepidemiology & therapeutic risk management* (pp. 22–23).

Berlin, J. A., Glasser, S. C., & Ellenberg, S. S. (2008). Adverse event detection in drug development: recommendations and obligations beyond phase 3. *American Journal of Public Health*, *8*, 1366–1371.

DuMouchel, W. (1999). Bayesian data dining in large frequency tables, with an application to the FDA spontaneous reporting system. *American Statistician*, *3*, 177–190.

DuMouchel, W., & Pregibon, D. (2001). Empirical Bayes screening for multi-item associations. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 67–76). New York: ACM.

Edwards, I. R., & Biriell, C. (1994). Harmonisation in pharmacovigilance. *Pharmacoepidemiology and Drug Safety*, *2*, 93–102.

Evans, S. J. W., Waller, P. C., & Davis, S. (2001). Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety*, *6*, 483–486.

Furberg, C. D., & Pitt, B. (2001). Withdrawal of cerivastatin from the world market. *Current Controlled Trials in Cardiovascular Medicine*, *5*, 205–207.

Garbe, E., Kloss, S., Suling, M., Pigeot, I., & Schneeweiss, S. (2012). High-dimensional versus conventional propensity scores in a comparative effectiveness study of coxibs and reduced upper gastrointestinal complications. *European Journal of Clinical Pharmacology*. doi:10.1007/s00228-012-1334-2

Grønbæk, H., Johnsen, S. P., Jepsen, P., Gislum, M., Vilstrup, H., Tage-Jensen, U., & Sørensen, H. T. (2008). Liver cirrhosis, other liver diseases, and risk of hospitalisation for intracerebral haemorrhage: a Danish population-based case-control study. *BMC Gastroenterology*, *1*. doi:10.1186/1471-230X-8-16

Hauben, M., Madigan, D., Gerrits, C. M., Walsh, L., & Van Puijenbroek, E. P. (2005). The role of data mining in pharmacovigilance. *Expert Opinion on Drug Safety*, *5*, 929–948.

Johnsen, S. P., Pedersen, L., Friis, S., Blot, W. J., McLaughlin, J. K., Olsen, J. H., & Sørensen, H. T. (2003). Nonaspirin nonsteroidal anti-inflammatory drugs and risk of hospitalization for intracerebral hemorrhage. *Stroke*, *2*, 387–391.

Mattle, H., Kohler, S., Huber, P., Rohner, M., & Steinsiepe, K. F. (1989). Anticoagulation-related intracranial extracerebral haemorrhage. *Journal of Neurology, Neurosurgery and Psychiatry*, *52*, 829–837.

Morris, J. A., & Gardner, M. J. (1988). Calculating confidence-intervals for relative risks (odds ratios) and standardized ratios and rates. *British Medical Journal*, *6632*, 1313–1316.

Olsen, M., Johansen, M. B., Christensen, S., & Sørensen, H. T. (2010). Use of vitamin K antagonists and risk of subarachnoid haemorrhage: a population-based case-control study. *European Journal of Internal Medicine*, *4*, 297–300.

Oracle (2011). *Oracle health sciences empirica signal*. http://www.oracle.com/us/industries/life-sciences/health-sciences-empirica-signal-364243.html. Cited in July 19, 2011

Pigeot, I., & Ahrens, W. (2008). Establishment of a pharmacoepidemiological database in Germany: methodological potential, scientific value and practical limitations. *Pharmacoepidemiology and Drug Safety*, *3*, 215–223.

Rassen, J. A., & Schneeweiss, S. (2012). Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. *Pharmacoepidemiology and Drug Safety*, *21*, 41–49.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *1*, 41–55.

Rothman, K. J., Lanes, S., & Sacks, S. T. (2004). The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiology and Drug Safety*, *8*, 519–523.

Schneeweiss, S., Rassen, J., Glynn, R. J., Avorn, J., Mogun, H., & Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, *4*, 512–522.

Schuemie, M. J. (2011). Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. *Pharmacoepidemiology and Drug Safety*, *3*, 292–299.

Shibata, A., & Hauben, M. (2011). Pharmacovigilance, signal detection and signal intelligence overview. In *Proceedings of the 14th international conference on information fusion (FUSION)* (pp. 1–7).

Sturgeon, J. D., Folsom, A. R., Longstreth, W. T., Shahar, E. Jr., Rosamond, W. D., & Cushman, M. (2007). Risk factors for intracerebral hemorrhage in a pooled prospective study. *Stroke*, *10*, 2718–2725.

Vandenbroucke, J. P., & Psaty, B. M. (2008). Benefits and risks of drug treatments: how to combine the best evidence on benefits with the best data about adverse effects. *JAMA*, *20*, 2417–2419.

Van Puijenbroek, E. P., Bate, A., Leufkens, H. G. M., Lindquist, M., Orre, R., & Egberts, A. C. G. (2002). A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and Drug Safety*, *1*, 3–10.

Vray, M., Hamelin, B., & Jaillon, P. (2005). The respective roles of controlled clinical trials and cohort monitoring studies in the pre- and postmarketing assessment of drugs. *Therapie*, *4*, 339.

Chapter 22
Meta-Analysis of Trials with Binary Outcomes

Jürgen Wellmann

22.1 Introduction

Clinical trials or observational epidemiological studies that investigate the health effects of a certain new treatment, a lifestyle factor, or an environmental condition, are often conducted in a similar manner by different teams of scientists in various places. In this way the uncertainties of single studies, and especially of small studies, can be tackled. When publications of these studies accumulate in the scientific literature, a systematic review is valuable that gathers, appraises, and summarizes this evidence. If the studies were conducted under comparable conditions and in nearly the same manner, and if they report their findings in terms of the same effect measure, their statistical results may be summarized quantitatively. Such an effort is called meta-analysis. Note that the contribution by Morgenthaler and Staudte in Chap. 19 also contains material highly relevant for meta-analyses.

The current chapter is concerned with statistical methods for the meta-analysis of studies that investigate the effect of a binary explanatory variable on a binary outcome. To be more concrete, this topic is discussed in terms of the meta-analysis of clinical trials that compare two treatments, say active treatment and placebo, and a clinical outcome. Of course the methods discussed here are applicable in other areas as well. The outcome usually is an unfavorable medical "event", like failure of medical care, worsening of symptoms, or even death. We concentrate on the odds ratio as the measure to compare the effect of active treatment versus placebo. Throughout the chapter, it will be assumed that no potential confounders need to be considered, as it is the case in randomized trials.

The results of such trials can be summarized in 2×2 tables of frequencies of "events" and "non-events" in both treatment groups. The numbers of events in each group are assumed to be binomially distributed. These frequencies can be analyzed by means of techniques that have been summarized, in the context of

J. Wellmann (✉)
Institute of Epidemiology and Social Medicine, University of Münster, 48149 Münster, Germany
e-mail: wellmann@uni-muenster.de

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_22, © Springer-Verlag Berlin Heidelberg 2013

meta-analyses, as “individual data methods” (Turner et al. 2000) or “bivariate meta-analysis” (Houwelingen et al. 2002). They range from long established methods for stratified 2 × 2 tables (Woolf 1955; Mantel and Haenszel 1959) to logistic regression with random effects. (See also the contribution by Suling, Weber and Pigeot, Chap. 21, for the analysis of stratified 2 × 2 contingency tables in pharmacoepidemiology.)

The term bivariate is justified from the point of view that the single studies constitute the “subjects” or units of analysis of the meta-analysis, whereby each studies supplies two observations, one for each treatment group. Each observation contains the (fixed) number of participants under study and the (random) number of events that occurred in the respective treatment group. These observations are uncorrelated across the trials, but may be correlated within the single trials.

On the other hand, “summary data methods” or “univariate meta-analyses” only require that one observation per trial is abstracted from the corresponding publications. These observations contain an estimate (here the odds ratio) and an appropriate measure of its variation (see Sect. 22.2.2 for details). A meta-analysis of such data usually involves the assumption that the logarithm of the odds ratio approximately follows a normal distribution. Furthermore, it is often assumed that the observed variances are fixed and known rather than random. This kind of analysis with its both variants as fixed effects or random effects meta-analysis (see Sect. 22.2.1) seems to be the classical approach.

There are some papers that investigate the statistical properties of the various methods sketched above. For example, Hartung and Knapp (2001) suggest a variant of the classical, univariate approach that accounts for the random variation of the observed measures of variation and compare this new variant with its classical predecessor by means of simulation. In another simulation Kuß and Gromann (2007), compare the two classics with the Mantel–Haenszel approach. These authors concentrate on tests for the hypotheses that the odds ratio equals one.

These papers concentrate on a few methods each. The purpose of the current chapter is to take a broader view and thus to give, firstly, an overview of the various univariate and bivariate methods for meta-analysis of trials with binary outcome and two treatment groups (Sect. 22.2). Secondly, some logistic regression approaches for the number of events in both treatment groups (Turner et al. 2000) are considered in more detail. An attempt is made to improve these methods by utilizing ‘sandwich-type’ estimators for the covariance matrix of the parameter estimates or a certain penalized likelihood. Finally, this broad range of methods is compared by means of simulations (Sect. 22.3), with an emphasis on estimation rather than testing. In the end, these results should help researchers to choose the most appropriate method for their meta-analysis (Sect. 22.4).

The emphasis on estimation (and confidence intervals) rather than testing in the current chapter is in line with the prevailing view in epidemiology and clinical research that knowing the magnitude of a health-related effect is more valuable than just knowing whether it is statistical significant (see, for example, Altman et al. 2000). The arguments in favour of estimation are similar to the motivation of Kulinskaya et al. (2008) who advocate a new measure of statistical evidence that is, amongst others, suitable for accumulating evidence of single studies in the framework of a meta-analysis. See the contribution by Morgenthaler and Staudte in Chap. 19 for a brief account of this approach. The current chapter is related to robust statistics insofar as empirical sandwich estimators for (co-)variance parameters are also considered, which depend less on distributional assumption than maximum likelihood estimators.

22.2 Model and Methods

22.2.1 A Logistic Gaussian Mixed Model

Let p_{ij} denote the (conditional) probability of the unfavorable medical event in the ith trial and treatment group j that pertains to n_{ij} subjects in the respective study group. For given p_{ij} the number of events Y_{ij} is then binomially distributed

$$Y_{ij} \sim \mathrm{B}(n_{ij}, p_{ij}), \quad i = 1, \ldots, k; \;\; j = 1, 2. \tag{22.1}$$

The probabilities p_{ij} need to be specified more closely. It is reasonable to allow for different levels of the response probability across studies, so that a model with trial-specific intercepts $\beta_0 + b_i$ is warranted. Furthermore, it is common to consider random treatment effects u_i, that are assumed to be realizations of independent random variables $U_i \sim N(0, \tau^2)$. Finally, the treatment group is specified by the fixed, observable variable x_{ij}. This variable is coded so that the fixed parameter θ is the overall log. odds ratio for the treatment effect. A logistic model, with link function $\mathrm{logit}(p) = \ln(p/(1-p))$, is now given by

$$\mathrm{logit}(p_{ij}) = \beta_0 + b_i + x_{ij}\theta + x_{ij}u_i, \quad i = 1, \ldots, k; \;\; j = 1, 2. \tag{22.2}$$

The trial-specific intercepts are the sum of some unknown parameter β_0 plus b_i, where the b_i can either be fixed parameters (with $b_k = 0$, say) or realizations of independent random variables $B_i \sim N(0, \sigma^2)$. One may assume that U_i and B_i are uncorrelated or that $\mathrm{Cov}(U_i, B_i) = \rho\sigma\tau$, see Turner et al. (2000, Eq. (3)).

One way to code treatment is to assign x_{ij} values of one and zero in the active treatment or in the placebo group, respectively. A more symmetric coding is $x_{i1} = 1/2$ for the active treatment and $x_{i2} = -1/2$ for the placebo group, see Turner et al. (2000). In both cases, θ is the unknown log. odds ratio which is to be estimated.

Fixed and Random Effects Meta-Analysis The "between trial variance" τ^2 may be constrained to equal zero. Then the u_i vanish and (22.2.1) is called a "fixed effects" (FE) model. In the context of meta-analyses, this term denotes a model that only contains a single, fixed parameter (here θ) for treatment effect across all studies. In a "random effects" (RE) model $\tau^2 \geq 0$ and thus a trial-specific deviation u_i from the overall effect θ is allowed.

In the RE model $\tau^2 = 0$ is explicitly allowed, as always in mixed models. Thus the FE model is a special case of the RE model. In principal, one has to decide beforehand what model is appropriate for the trials that are to be analyzed. The FE model is justified if one can be sure that identical treatment regimes have been tested in all trials, and thus the treatment effect should be the same in all studies. In practice one is more often confronted with treatment regimes that are comparable, but not identical, and thus give rise to the RE model.

One may be tempted to have a look at the data to assist in the choice of the model, but the effect of this data snooping on the statistical properties of the subsequent meta-analysis may be hard to state precisely. However, some meta-analysis procedures for the RE model have a kind of built-in model choice insofar as they are identical to a corresponding FE method as soon as their estimate for τ^2 becomes zero. Hartung and Knapp (2001) discuss this issue in the context of the truncated estimator of DerSimonian and Laird (1986), see (22.9) below, which equals zero with positive probability.

Subject Specific Versus Population Averaged Approaches Note that in the mixed effects versions of model (22.2.1), with $\tau^2 > 0$ and/or b_i random, p_{ij} is a conditional mean, $p_{ij} = \mathrm{E}(Y_{ij} \mid u_i, b_i)$. Mixed models are sometimes called subject specific models, since they make specific assumption on the single subjects under study (Zeger et al. 1988). The subjects of a meta-analysis are the single studies. Here, model (22.2.1) implies for the (conditional) log. odds ratio in trial i

$$\ln(\mathrm{OR})_i = \mathrm{logit}(p_{i1}) - \mathrm{logit}(p_{i2}) = \theta + u_i, \quad i = 1, \ldots, k, \tag{22.3}$$

for both codings for x_{ij}. Some methods for meta-analysis do not involve random effects and adopt an unconditional, 'population averaged' point of view. From this perspective, one is interested in the unconditional mean $\pi_{ij} = \mathrm{E}(\mathrm{E}(Y_{ij} \mid u_i, b_i))$ and the unconditional log. odds ratio $\mathrm{logit}(\pi_{i1}) - \mathrm{logit}(\pi_{i2})$. Following Zeger et al. (1988, p. 1054), one can derive an approximate relations between p_{ij}, π_{ij} and the fixed parameters of model (22.2.1). For b_i fixed and $x_{ij} = \pm 1/2$ one obtains

$$\mathrm{logit}(\pi_{ij}) \approx a\big(\tau^2\big)(\beta_0 + b_i \pm \theta/2) \tag{22.4}$$

and thus an unconditional log. odds ratio of $a(\tau^2)\theta$, where $a(\tau^2) = (c^2\tau^2/4 + 1)^{-1/2}$ with $c = 16\sqrt{3}/(15\pi) \approx 0.588$. Note that $a(\tau^2)$ is close to one as long as τ^2 is not large.

22.2.2 *The Univariate Model for Meta-Analysis*

Let $\hat{\theta}_i$ be the estimator for the conditional log. odds ratio (22.3) from the ith trial. It is assumed that this estimator in this specific trial, given u_i, is at least approximately normally distributed with variance σ_i^2, that is $\hat{\theta}_i \mid u_i \sim N(\theta + u_i, \sigma_i^2)$. Since $U_i \sim N(0, \tau^2)$ one obtains the usual assumption for the unconditional distribution of the estimators

$$\hat{\theta}_i \sim N\big(\theta, \tau^2 + \sigma_i^2\big), \quad i = 1, \ldots, k. \tag{22.5}$$

As in the previous section, one may choose between the FE approach, where $\tau^2 = 0$ is assumed, or the RE approach with $\tau^2 \geq 0$.

In a meta-analysis, $\hat{\theta}_i$ and an estimate of its variance, $\widehat{\sigma_i^2}$, may be obtainable from the published papers. These statistics are usually not directly available but may be easily derived from odds ratios and the corresponding confidence intervals. If the frequencies Y_{ij} and n_{ij} are available (which would enable a bivariate analysis as well), the desired "univariate" estimates can also easily be computed. The usual estimate of the odds ratio from a 2 × 2-table encounters problems if one of the four cells of the table contains a frequency of zero. A simple remedy for these sampling zeros is to add 0.5 to all four cells, and thus to estimate the odds ratio and its logarithm by

$$\widehat{\mathrm{OR}_i} = \frac{Y_{i1}+0.5}{n_{i1}-Y_{i1}+0.5}\,\frac{n_{i2}-Y_{i2}+0.5}{Y_{i2}+0.5} \quad \text{and} \quad \hat{\theta}_i = \ln(\widehat{\mathrm{OR}_i}). \tag{22.6}$$

The variance of $\hat{\theta}_i$ is estimated by

$$\widehat{\sigma_i^2} = \frac{1}{Y_{i1}+0.5} + \frac{1}{n_{i1}-Y_{i1}+0.5} + \frac{1}{Y_{i2}+0.5} + \frac{1}{n_{i2}-Y_{i2}+0.5}, \qquad i = 1,\dots,k. \tag{22.7}$$

Some authors suggest to always add 0.5 to all counts in 2 × 2 tables under study while others suggest to apply this remedy only to those 2 × 2 tables where sampling zeros occurred.

22.2.3 Statistical Methods for Meta-Analysis

The methods listed in this section will be compared by simulation.

Univariate Meta-Analysis The log. odds ratios (22.6) can be used to estimate the common odds ratio θ and appropriate confidence intervals. See, for example, Hartung and Knapp (2001) for a nice derivation of these classical meta-analysis methods. They amount to estimating θ by a weighted mean of the published estimates. These weights contain estimates (22.7) of the within-trial variance, and in the RE approach also the between-trial variance τ^2. Classical approaches for meta-analysis neglect the random variation of these estimates, while Hartung and Knapp do account for it.

- For the FE approach, weights $\hat{v}_i = 1/\hat{\sigma}_i^2$ are needed together with $\hat{v} = \sum_{i=1}^k \hat{v}_i$. The estimator for the common log. odds and its $(1-\alpha)$ confidence interval is

$$\hat{\theta}_{\mathrm{FE}} = \frac{1}{\hat{v}} \sum_{i=1}^{k} \hat{v}_i \hat{\theta}_i, \qquad \hat{\theta}_{\mathrm{FE}} \pm q_{1-\alpha/2}/\sqrt{\hat{v}}, \tag{22.8}$$

  where q_γ denotes the γ-quantile of the standard normal distribution, see for example Hartung and Knapp (2001). This method has already been suggested by Woolf (1955) for a common odds ratio in $k \times 2 \times 2$ tables.
- The estimator in the RE model is build analogous to (22.8) with $\hat{v}_i$ replaced by $\hat{w}_i = (\hat{\tau}^2 + \hat{\sigma}_i^2)^{-1}$ and $\hat{w} = \sum_{i=1}^k \hat{w}_i$, where $\hat{\tau}^2$ estimates the between-trial variance τ^2. A frequently used estimator is the truncated version of the moment estimator suggested by DerSimonian and Laird (1986)

$$\hat{\tau}^2 = \frac{\max\{0, \sum_{i=1}^k \hat{v}_i (\hat{\theta}_i - \hat{\theta}_{\mathrm{FE}})^2 - (k-1)\}}{\hat{v} - \sum_{i=1}^k \hat{v}_i^2/\hat{v}}. \tag{22.9}$$

The classical approach in the RE model is to use

$$\hat{\theta}_{\mathrm{RE}} = \frac{1}{\hat{w}} \sum_{i=1}^{k} \hat{w}_i \hat{\theta}_i, \qquad \hat{\theta}_{\mathrm{RE}} \pm q_{1-\alpha/2}/\sqrt{\hat{w}}. \tag{22.10}$$

- Following Hartung and Knapp (2001) one may construct a third kind of confidence interval:

$$\hat{\theta}_{\mathrm{RE}} \pm t_{k-1;1-\alpha/2}\sqrt{\hat{Q}} \quad \text{with } \hat{Q} = \frac{1}{k-1} \sum_{i=1}^{k} \frac{\hat{w}_i}{\hat{w}} (\hat{\theta}_i - \hat{\theta}_{\mathrm{RE}})^2, \tag{22.11}$$

  where $t_{\nu;\gamma}$ denotes the γ-quantile of the t-distribution with ν degrees of freedom.

A preliminary simulation suggested that always adding 0.5 to all cells of the $k \times 2 \times 2$ table yields a slightly larger bias for the estimators of θ. Therefore, in the following 0.5 is only added to those 2×2 tables where sampling zeros occurred.

Bivariate Meta-Analysis If the quantities Y_{ij} and n_{ij} of the binomial model (22.2.1) are available, one may directly estimate the parameters of this model following the usual approaches for logistic regression with or without random effects. The following variants are considered.

- Logistic regression with fixed, trial-specific intercepts and an overall parameter for treatment; no random treatment effect ($\tau^2 = 0$, FE model). Since a preliminary simulation revealed many cases of complete separation, logistic regression with Firth's penalized maximum likelihood (Heinze and Schemper 2002; Firth 1993) was employed too.
- Logistic regression for an FE model with random trial-specific intercepts.
- Mixed effects logistic regression with random trial-specific intercepts, random treatment effect ($\tau^2 \geq 0$, RE model), and a fixed overall parameter for treatment. The two random effects are assumed to be independent.
- As above, but with correlated random effects.

In our simulation always the coding $x_{ij} = \pm 1/2$ was used. The latter three approaches involve random effects and are also computed with the sandwich-type covariance estimator of Mancl and DeRouen (2001).

- An intriguing alternative in the sense of an FE model is the estimator of Mantel and Haenszel (1959) for a common odds ratio in a $k \times 2 \times 2$ contingency table

$$\widehat{\mathrm{OR}}_{\mathrm{MH}} = \left\{ \sum_{i=1}^{k} \frac{Y_{i1}(n_{i2} - Y_{i2})}{n_{i1} + n_{i2}} \right\} \left\{ \sum_{i=1}^{k} \frac{Y_{i2}(n_{i1} - Y_{i1})}{n_{i1} + n_{i2}} \right\}^{-1}. \tag{22.12}$$

  It is classified here into the bivariate methods since one needs the results from both treatment groups to calculate the estimator. A $(1 - \alpha)$ confidence interval is constructed as $\ln(\widehat{\mathrm{OR}}_{\mathrm{MH}}) \pm q_{1-\alpha/2}\hat{\sigma}_{\mathrm{MH}}$, where $\hat{\sigma}^2_{\mathrm{MH}}$ denotes the estimate of Robins et al. (1986) of the variance of the logarithm of $\widehat{\mathrm{OR}}_{\mathrm{MH}}$.

22.2.4 Simulation

Data were simulated according to the binomial model in Sect. 22.2.1, similar to Kuß and Gromann (2007). A simulated meta-analysis consists of k trials, $k = 5, 10, 20$, where each trial supplies two numbers Y_{i1} and Y_{i2} from a binomial distribution $B(n_{ij}, p_{ij})$. The number of participants is set to the same value $n_{11} = \cdots = n_{k1} = n_{21} = \cdots = n_{k2} = 20, 50$ across all trials and both treatment groups.

The probabilities p_{ij} of (22.2) involve $\theta = 0, \ln(2/3), \ln(1/3)$, and thus give rise to odds ratios $1, 2/3, 1/3$. Random treatment effects u_i were generated from a normal distribution with mean 0 and variance τ^2 with $\tau = 0, 0.05, 0.1, 0.5$. Treatment is coded by $x_{ij} = \pm 1/2$. The trial-specific intercepts are fixed with $b_1 = \cdots = b_k \equiv 0$ and $\beta_0 = -\ln(9), 0$. Define p such that $\text{logit}(p) = \beta_0$, that is to say $p = 0.1$ or $p = 0.5$, respectively. These are the probabilities p_{ij} in the null model with $\theta = 0$ and $\tau^2 = 0$, respectively. In the other situations, the p_{ij} are distributed below and above these values.

In summary, $3(k) \times 2(n_{ij}) \times 2(\beta_0) \times 3(\theta) \times 4(\tau) = 144$ different situations are simulated. Each time, 2,000 simulation runs were performed.

The simulated data contain the frequencies Y_{ij} and n_{ij} that are needed for the bivariate methods. The estimates needed as input for the univariate methods are computed thereof as in (22.6) and (22.7).

For each of the 144 simulated situations the three univariate methods, the Mantel–Haenszel approach, and the variants of logistic regression mentioned above are considered. Each method yields an estimate of the common log. odds ratio plus a 95 % confidence interval and an estimate for τ^2, if appropriate.

For each method the percentage of successful simulation runs, i.e., runs without numerical problems, was recorded. Especially the regression methods did not always arrive at a solution due to convergence problems. For all successful runs, the average of the estimators for θ was computed, as well as the percentage of simulation runs where the confidence interval for θ contains the true θ.

All computations were carried out in SAS® version 9.2 (SAS Institute Inc., Cary, NC, USA), using the random number generators RANNOR and RANBIN and the procedures FREQ, LOGISTIC and GLIMMIX. The SUMMARY procedure was used to compute the ingredients for $\hat{\theta}_{\text{RE}}$ and the corresponding confidence intervals.

22.3 Results

22.3.1 Computational Issues

For each situation, 2,000 simulation runs were performed and the percentage of successful runs was recorded. Table 22.1 lists the mean and the minimum of these percentages over the 144 situations under study. The univariate FE (22.8) and RE approach (22.10) did not provide any computational problems. Neither does the alternative confidence interval (22.11).

Table 22.1 Overview of simulation results. Worst and average results with regard to computational problems and bias of estimates for treatment effect θ

Procedure			Computation		Bias		
Assumptions for effect of			Successful runs [%]		Geometric mean		
trial	treatment	abbr.[a]	min	mean	min	mean	max
Univariate							
	FE (22.8)	Uni-FE	100.0	100.0	0.98	1.03	1.20
	RE (22.10)[b]	Uni-RE	100.0	100.0	0.96	1.02	1.20
Bivariate							
Mantel–Haenszel		BiF-MH	99.6	100.0	0.95	1.00	1.03
logistic regression							
fixed	FE	BiF-FE	75.6	97.0	0.92	0.99	1.01
	FE+Firth	+Firth	100.0	100.0	0.97	1.00	1.01
	RE[c]	BiF-RE	74.9	96.8	0.90	0.98	1.01
random	RE, Cor = 0[c]	BiR-REind	97.3	99.3	0.92	0.99	1.01
	RE, Cor = ρ[c]	BiR-REcor	84.4	92.3	0.93	0.99	1.01

[a] Abbreviation for use in figure legends

[b] The univariate RE approach with confidence interval (22.11) and

[c] the logistic RE models with sandwich estimators yield the same results as their counterparts

Cor = correlation between trial-specific intercepts and random treatment effects

The Mantel–Haenszel approach encountered problems in those few simulated meta-analyses where no event occurred in all groups with active treatment. In the worst case, this happened in 8 out of 2,000 runs, which yields the 99.6 % of successful runs listed in Table 22.1.

The iterative solution of the logistic regression problems quite often fail to converge in some situations. This is especially true for the models with fixed trial effects, unless Firth's penalized likelihood is employed. With this approach, which was only available for the FE model, the algorithm always converged. The convergence problems of the ordinary maximum likelihood estimation almost exclusively occurred in the meta-analyses with rare events ($p = 0.1$) and only 20 participants per treatment arm. The frequency of convergence problems increase from 6.8 % in meta-analyses with $k = 5$ trials to about a quarter with $k = 20$ (Table 22.2).

Logistic regression with random trial effects suffers less from computational problems than the model with fixed trial effects and the standard likelihood. Among the two logistic regression methods of this type the version that allows for a correlation between the trial and the random treatment effects more often fails to converge. In some situations, only about 85 % of simulation runs terminate successfully. The

Table 22.2 Percentage of successful simulation runs in bivariate logistic regression

Assumption for type of effects		k	$n_{ij} = 20$				$n_{ij} = 50$			
			$p = 0.1$		$p = 0.5$		$p = 0.1$		$p = 0.5$	
trial	treat.		min	max	min	max	min	max	min	max
fixed	FE	5	93.2	96.0	100.0	100.0	100.0	100.0	100.0	100.0
		10	87.0	93.3	100.0	100.0	100.0	100.0	100.0	100.0
		20	75.6	85.7	100.0	100.0	100.0	100.0	100.0	100.0
fixed	RE	5	93.2	96.0	99.9	100.0	100.0	100.0	99.9	100.0
		10	86.4	92.7	99.8	100.0	99.9	100.0	99.0	100.0
		20	74.9	85.0	99.6	100.0	99.9	100.0	98.8	100.0
random Cor = 0	RE	5	98.0	99.9	99.8	100.0	98.8	99.9	99.4	100.0
		10	98.1	99.9	99.1	99.7	98.8	99.6	98.9	99.9
		20	97.3	99.3	98.5	99.2	97.9	99.7	98.3	99.9
random Cor = ρ	RE	5	88.9	91.8	91.0	96.1	90.8	95.7	90.1	96.5
		10	86.5	90.5	90.9	95.2	90.8	95.0	89.1	96.5
		20	85.0	90.0	88.7	92.9	89.3	93.5	84.4	95.4

Cor = correlation between trial-specific intercepts and random treatment effects

regression with uncorrelated random effects fails to converge mostly in one or two percent of all runs (Table 22.2).

The sandwich estimators, which have also been tested in the logistic RE models, have no effect on convergence problems.

22.3.2 Bias

We simulated the mean of the estimators for the log. odds ratio as the average of the respective estimates, and the bias as this average minus the true θ. For the overview in Table 22.1 we take the antilogarithm of the simulated bias and thus present the geometric mean of the "multiplicative bias" $\widehat{\text{OR}}/\exp(\theta)$.

Logistic regression with fixed trial effects and without random treatment effects yields values close to one, that is to say there is nearly no bias. This is especially true for the variant with Firth's likelihood. The results for the Mantel–Haenszel odds ratios are also quite close to one.

In some situations, the univariate procedures tend to overestimate the odds ratio. A closer look at the simulation results reveals that this effect occurs if OR < 1 and the outcomes are rare ($p = 0.1 \Leftrightarrow \beta_0 = -\ln(9)$), see Fig. 22.1. Overall, the RE estimator is on average slightly closer to θ than its FE counterpart.

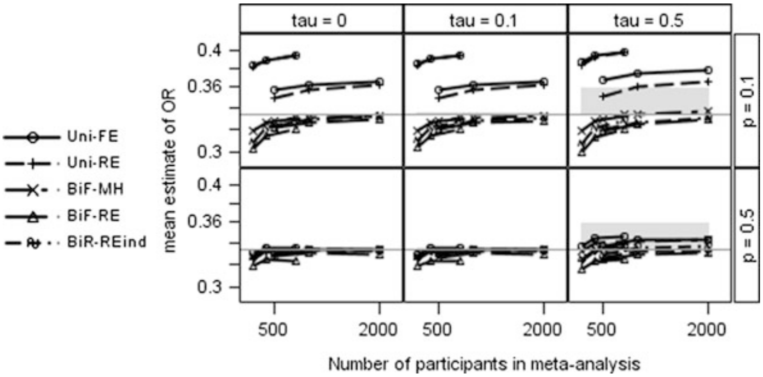


Fig. 22.1 Simulated means of estimates for the log. odds ratio for selected procedures; true $\mathrm{OR} = 1/3$

The logistic regression procedures with random effects are prone to underestimation. Figure 22.1 demonstrates that this phenomenon especially occurs in meta-analysis which comprise a smaller number of participants. No approach is clearly superior to the others.

More details are given in Fig. 22.1. This graphic depicts the simulation results for the estimates of θ in terms of the mean of the simulated estimates for θ. These means are plotted against the number of subjects in the simulated meta-analyses, $k \times 2 \times n_{ij}$. Different symbols are used to distinguish the methods. The symbols for each methods are connected by lines, but each line is broken into two parts to distinguish the results for $n_{ij} = 20$ from those for $n_{ij} = 50$ subjects per treatment group. Results for $\tau = 0.05$ are in line with the other results and are omitted in the graphics. A horizontal line indicates the true $\mathrm{OR} = \exp(\theta)$. A gray bar extends from this line to $\exp(a(\tau^2)\theta)$, cf. (22.4), to indicate the corresponding population averaged odds ratio.

The simulation results for $\mathrm{OR} = 2/3$ show a pattern very similar to Fig. 22.1 but with much smaller bias. For $\mathrm{OR} = 1$ the pattern of the results is different, but the bias is even smaller, see Figs. 22.2 and 22.3.

22.3.3 Coverage

95 % confidence intervals for the log. odds ratio θ were computed for each method. We simulated the percentage of these intervals that contain the true θ.

The results for these coverage probabilities are mainly influenced by the between-trial variance τ^2, with one exception. In meta-analysis with rare events ($p = 0.1 \Leftrightarrow \beta_0 = -\ln(9)$) and only n_{ij} participants per treatment group, the coverage probability of the univariate RE approach with the confidence interval based on the t-distribution (22.11) drops with increasing number of trials to about 90 % for $k = 20$.

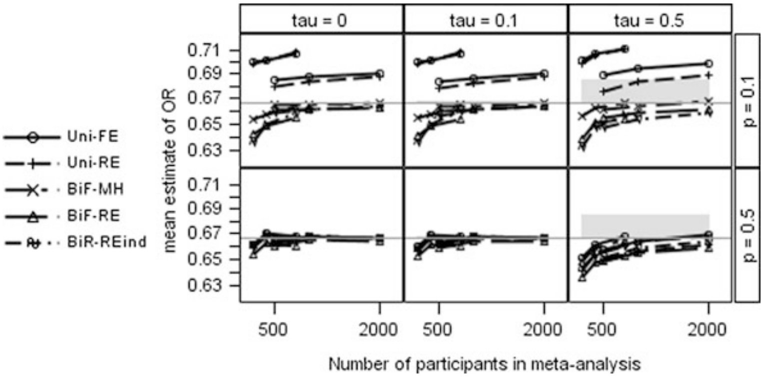


Fig. 22.2 Simulated means of estimates for the log. odds ratio for selected procedures; true OR = 2/3

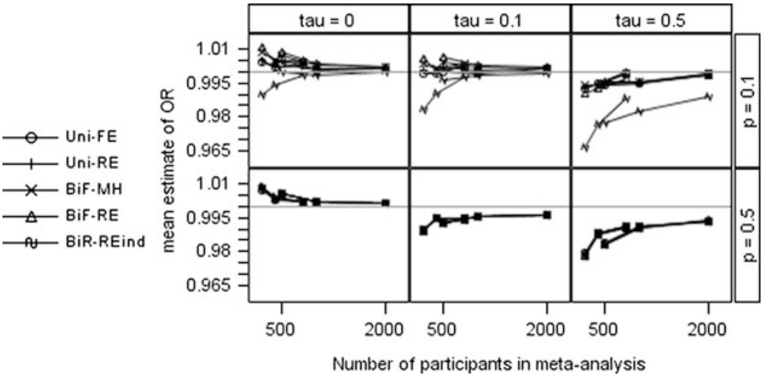


Fig. 22.3 Simulated means of estimates for the log. odds ratio for selected procedures; true OR = 1

Apart from this finding, we observe that all procedures achieve a coverage of about 95 % or more for $\tau^2 = 0$. With increasing τ^2 the FE methods yield confidence intervals that contain θ less often (Fig. 22.4). Results for coverage probabilities are presented for OR = 2/3. The results for OR = 1 and OR = 1/3 show essentially the same patterns.

The same is true for three RE methods, namely the univariate RE procedure with confidence interval (22.10), see Fig. 22.5, and the two logistic regression procedures with random trial effects and standard confidence intervals (Fig. 22.6). Their coverage probability is especially low for $\tau = 0.5$. The situation becomes worse with increasing number of participants and is most pronounced in meta-analyses of frequent events ($p = 0.5$).

All other RE methods comply with the 95 % confidence level. This implies that the confidence interval (22.11), see again Fig. 22.5, and the sandwich esti-

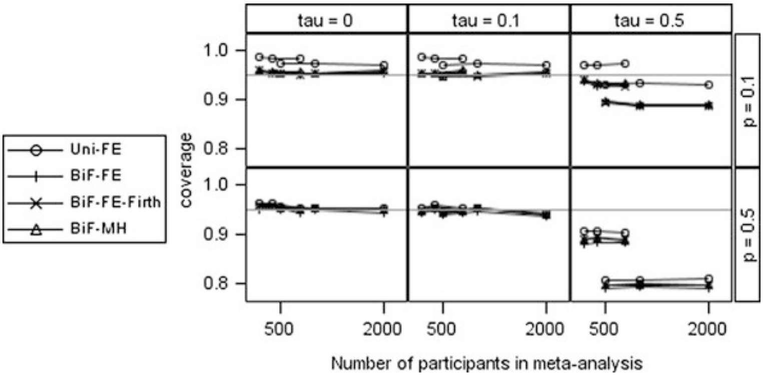


Fig. 22.4 Simulated coverage probabilities for the confidence interval for the log. odds ratio; true OR = 2/3, FE methods

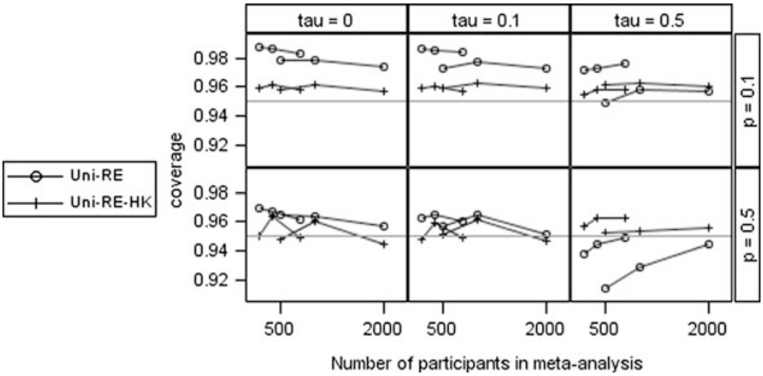


Fig. 22.5 Simulated coverage probabilities for the confidence interval for the log. odds ratio; true OR = 2/3, univariate RE methods with 'normal' confidence intervals (22.10) and those suggested by Hartung and Knapp (22.11)

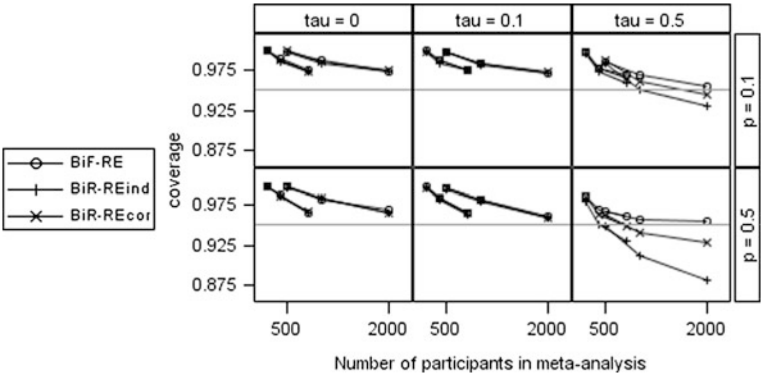


Fig. 22.6 Simulated coverage probabilities for the confidence interval for the log. odds ratio; true OR = 2/3, bivariate RE methods with standard confidence intervals

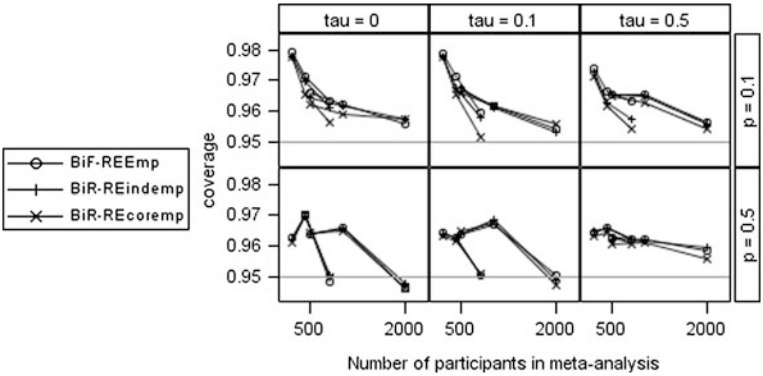


Fig. 22.7 Simulated coverage probabilities for the confidence interval for the log. odds ratio; true OR = 2/3, bivariate RE methods, confidence intervals based on empirical sandwich estimates

mators (Fig. 22.7), respectively, improve the confidence intervals of the respective approaches.

22.4 Conclusions

Introductory texts on meta-analysis often start with methods that require one statistic per trial, plus an estimate of its variation, and an assumption that this statistic is approximately normally distributed. This approach is reasonable and may often be the only feasibly option as long as no other information is available. However, if studies are to be analyzed that compare two groups of subjects, data on both groups may be available and offer the opportunity for a bivariate meta-analysis (Houwelingen et al. 2002). Especially in clinical trials with a binary endpoint the results of a single trial may be presented as number of "events" and number of subjects in both groups. Meta-analysis of such data amounts to the analysis of a $k \times 2 \times 2$ table. Appropriate methods, as for example the Mantel–Haenszel odds ratios, have been known for decades and seem to be a more natural approach than methods that rely on a normal distribution of the logarithm of the odds ratio. And if random treatment effects are warranted, as is often the case in meta-analyses, logistic regression offers the desired options (Turner et al. 2000).

The univariate approaches considered here involve adding a small increment to the cell frequencies if sampling zeros occur. Without this "continuity correction" there would have been many instances where the estimators could not have been computed. However, this measure has some drawbacks, see, for example, the discussion in Rücker et al. (2009).

The Mantel–Haenszel Odds Ratio and the corresponding confidence interval is computable even with many sampling zeros, unless certain patterns of empty cells in each of the 2×2-tables occur. Thus one may even regard numerical problems of the Mantel–Haenszel approach as an indication of remarkable results rather than a drawback of the method.

Logistic regression with fixed trial-specific intercepts involves a large number of fixed parameters, which makes it prone to problems of complete separation of the data, a situation where the maximum likelihood estimator of the logistic regression does not exist. Use of the penalized likelihood suggested by Firth (1993) avoids this problem, as was first observed by Heinze and Schemper (2002).

Models with random intercepts contain much less fixed parameters and encounter less numerical problems than the standard logistic regression with fixed intercepts. In a concrete meta-analysis, these problems may be overcome by careful choice of starting values for the estimates or different numerical techniques. These options have not been tested in the current simulation.

In summary, the Mantel–Haenszel approach and the logistic regression with only fixed effects and Firth's likelihood are remarkably stable from a computational point of view, the numerical stability of the univariate approaches come at the cost of a questionable continuity correction, and the convergence problems of the logistic regression models with random effects should not deter one from using this methods. It would have been interesting to test an approach similar to Firth's likelihood in the mixed effects models, but this option was not supported by the software used here and is thus beyond the scope of the current work.

The bias of the estimators for the log. odds ratio are not too severe, given the fact that the worst results occurred in the rather extreme situation with random treatment effects with a standard deviation of $\tau = 0.5$, which is quite pronounced. But anyhow this finding is an argument against the use of the univariate methods in which these results were observed.

The Mantel–Haenszel Odds Ratios are, on average, very close to the true odds ratio. The logistic regression with fixed effects and Firth's likelihood performs even better. This is in accordance with Firth's intention who suggested the penalized likelihood in order to avoid the (asymptotic) bias of the maximum likelihood estimator.

It is not surprising that the FE methods, that are not suited to deal with random treatment effects, do not perform well if they are applied to data that were generated under a model with such random effects. We note that the confidence interval (22.11) improves the univariate RE approach, as already established by Hartung and Knapp (2001) (with the one exception described above) and that sandwich estimators improve the confidence intervals for the bivariate logistic regression models with random effects.

The current simulation suggests to use bivariate methods whenever the necessary information can be abstracted from the papers under study. The Mantel–Haenszel Odds Ratio and logistic regression with Firth's likelihood are good alternatives for a FE meta-analysis. If random effects are warranted one may use one of the logistic regression models (Turner et al. 2000), whereby sandwich estimates (Mancl and DeRouen 2001) improve the confidence intervals. If the trials that are to be analyzed only provide odds ratios and their standard errors, one should take into account that the latter are estimates and not the true values (Hartung and Knapp 2001).

References

Altman, D. G., Machin, D., Bryant, T. N., & Gardner, M. J. (2000). *Statistics with confidence. BMJ books* (2nd ed.).

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*, 27–38.

Hartung, J., & Knapp, G. (2001). A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, *20*, 3875–3889.

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, *21*, 2409–2419.

Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Tutorial in biostatistics: advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, *21*, 589–624.

Kulinskaya, E., Morgenthaler, S., & Staudte, R. G. (2008). *Meta analysis: a guide to calibrating and combining statistical evidence*. New York: Wiley.

Kuß, O., & Gromann, C. (2007). An exact test for meta-analysis with binary endpoints. *Methods of Information in Medicine*, *46*, 662–668.

Mancl, L. A., & DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics*, *57*, 126–134.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748.

Robins, J., Breslow, N., & Greenland, S. (1986). Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, *42*, 311–323.

Rücker, G., Schwarzer, G., Carpenter, J., & Olkin, I. (2009). Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine*, *28*, 721–738.

Turner, R. M., Omar, R. Z., Yang, M., Goldstein, H., & Thompson, S. G. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, *19*, 3417–3432.

Woolf, B. (1955). On estimating the relationship between blood group and disease. *Annals of Human Genetics*, *19*, 251–253.

Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimation equation approach. *Biometrics*, *44*, 1049–1060.